Universidad Autónoma de Madrid

Facultad de Ciencias

Departamento de Biología Molecular

Tesis Doctoral

The Human Phylome - A large scale phylogenetic study on the human genome evolution

Jaime Huerta Cepas

Septiembre - 2008

Director:
Dr. Joaquín Dopazo Blázquez

Co-Director:
Dr. Juan Antonio Gabaldón Estevan

Contents

Part I

Introduction

Chapter 1

Molecular phylogenies and Bioinformatics

1.1 Sequences

Sequences are central to molecular biology. They represent the ordered chain of elements that form a macromolecule and that, in the end, confer them their specific functions and properties. Such elements are typically nucleotides (basic components of nucleic acids) or amino acids (in peptides). Although proteins and genes are known since the eighteenth and nineteenth centuries, respectively, the birth of molecular biology as a field is not formally set until the 1950s, coinciding with the elucidation of the DNA structure (Watson and Crick, 1953), the first confirmation of the Central Dogma (Meselson and Stahl, 1958) and the discovery of the genetic code (a work started by Severo Ochoa and culminated by Marshall Warren Nirenberg, Robert W. Holley and Hargobind Khorana).

From then on, molecular biologists have learned to characterize, isolate and manipulate macromolecules in order to understand the cell's behavior. More recently, the history of molecular biology has been influenced by another key finding. In 1975, Frederick Sanger and collaborators presented a novel method to sequence DNA (Sanger and Coulson, 1975), which revolutionized, not only the field of molecular biology, but also many other areas in biology. Sanger's technique, named the chain terminator method (Sanger *et al.*, 1977), provided a relatively rapid method to determine the exact sequence of nucleotides from a DNA molecule. In fact, only two years later from the publication of the method, Sanger published the first sequence from a complete genome, that from the Phage (Fi)-X174 (Sanger *et al.*, 1977). Since then, the number of known sequences have been increasing continuously, providing a new extensive and valuable source of biological information. The importance of DNA sequencing techniques can be symbolized by the fact that Frederick Sanger was laureated in 1980, with his second Nobel Price.

1.2 Molecular phylogenies

Phylogenetics was, undoubtedly, one of the fields that more directly received the influence of the DNA sequencing revolution. For years, the study of the evolutionary relatedness among species was only possible by the careful comparison of those morphological characters that were considered as homologous among organisms. With the increasing availability of DNA and protein sequences phylogenetic surveys were allowed to move from phenotypes to genotypes, and use each sequence position as a comparable character. However, the use of such potential required the development of new phylogenetic methods and the use of computers. While morphological analyses did not usually exceed some tens of characters, sequence analyses required the comparison of hundreds or thousands of residues (one by each nucleotide or amino acid that compose a sequence). This made molecular phylogenetics to become highly dependent on computational approaches, thus contributing to the early development of the bioinformatics field. Indeed, one of the first computational analyses performed on biological sequences was a phylogenetic analysis (Zuckerkandl and Pauling, 1965).

Since those days, and over all the second second half of the twentieth century, phylogenetics and computational biology have stayed closely related. Many bioinformatics methods have been developed and fine tuned to perform reliable phylogenetic analysis on molecular sequences. Examples of these are the algorithms to build sequences alignment (Needleman and Wunsch, 1970; Smith and Waterman, 1981); the programs to perform phylogenetic reconstruction on sequences , or the definition of matrices to model the evolution of amino acids (Jones *et al.*, 1992; Henikoff and Henikoff, 1992; Müller and Vingron, 2000; Whelan and Goldman, 2001).

Nowadays, molecular phylogenies are recognized as a very valuable source of information, not only for taxonomists but also for geneticists and molecular biologists. For instance, the use of mitochondrial sequences has been crucial for deciphering many species relationships , co-evolution studies have been applied to the prediction of protein interactions (Pazos and Valencia, 2001; Ramani and Marcotte, 2003), and phylogenetic trees have been used for predicting the function of yet uncharacterized genes (Eisen, 1998; Gabaldón and Huynen, 2004)

1.2 Sequence homology

A phylogenetic analysis only makes sense in the context of sequences that are evolutionarily related. Such evolutionary relatedness is formally expressed through the concept of homology, a term coined by Owen much earlier than the advent of the molecular biology (Owen, 1848). As Owen originally defined in his anatomic studies, homology represents the condition of sharing a common ancestor. Hence, if we extend the term to modern biology, two sequences are considered homologous if they originated from the same ancestral sequence.

Establishing the correct homology relationship is extremely important in phy-

logenetics, since all the subsequent analyses and interpretations will rely on this first assumption. But, how can we identify homology? Since it represents an abstract concept that cannot be measured experimentally (we would need to know ancestral sequences), the only way to infer homology is through indirect evidence. The most widely used indicator of homology is sequence similarity. Strictly speaking, similarity can perfectly be the result of independent evolutionary convergence. The effect causing such convergence is known as homoplasy, and the resulting similarity is named analogy rather than homology. Fortunately for bioinformaticians, though, analogy cases seem to be more frequent among morphological characters than at genotypic levels. Although it is true that certain levels of sequence analogy can be found between short regions (such as protein domains), it is considered a very improbable process at the genic level (Fitch, 2000). For this reason, high (significant) levels of sequence similarity are usually taken as reliable evidence of homology. Of course, this procedure does not imply a synonymous relationship between homology and similarity. This fact is sometimes disregarded, and it is not rare to find references to homology when they actually refer to similarity: e.g 15% homology, or high homology. The misunderstanding resides in the fact that, while similarity is a concept that can be measured and experimentally tested, homology does only represents a conjecture about an absolute condition. Subsequent statistical analyses on the sequences phylogeny (such as bootstrap or posterior probability) may provide stronger support for the homology assumption.

1.2.1 Orthology and Paralogy

Homology relationships between sequences can be further sub-divided into orthology and paralogy relationships, which were precisely defined by Walter Fitch in 1970 (Fitch, 1970).

Two (or more) sequences are considered orthologous if they diverged from the same ancestral sequence through a speciation event (Figure 1.1). In other words, they are the result of a common sequence evolving separately after the differentiation of the ancestral organism into two new species. In contrast, paralogs are homologous sequences that originated from a duplication event (Figure 1.1). Because a duplication can only occur within a single genome, paralogs necessarily started to diverge within the same organism. Nevertheless, they may, later on, be inhered by the new lineages originated by the subsequent speciation events.

An important issue derived from the evolutionary definition of orthology and paralogy is that relationships are not limited to a couple of sequences. Actually, several genes can be considered, at the same time, as co-orthologs or co-paralogs to another set of genes. This idea is usually referred in the literature as one-to-one, one-to-many and many-to-many relationships.

Given that paralogs, but not orthologs, may produce states of genetic redundancy within a genome, it is assumed that they are more likely to diverge toward new functions. In contrast, orthologs are expected to conserve equivalent functions across species (Fitch, 1970). Although this can only be taken as a general

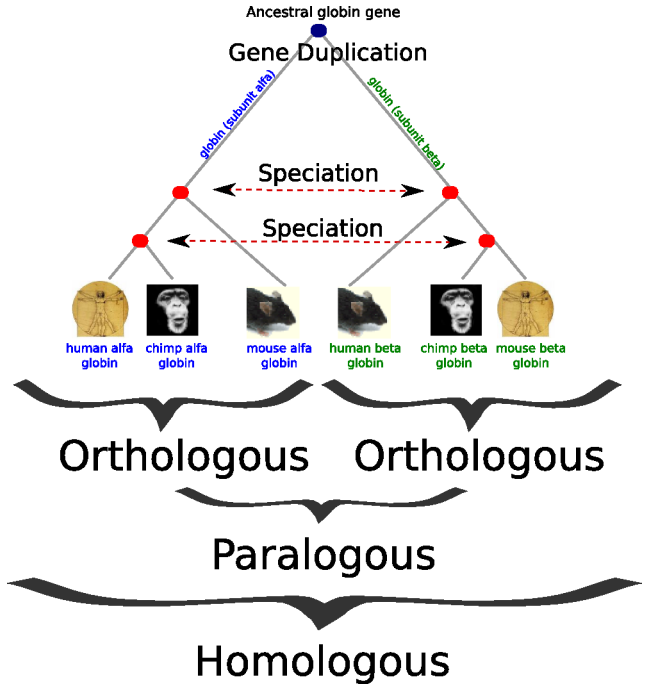


Figure 1.1: Schematic representation the orthology and paralogy relationships among a group of homologous sequences.

trend, it has deep implications for the transfer of functional information across organisms. For instance, the establishment of equivalences among genes from different genomes (orthologies) is a pre-requisite for comparing genomics data, something that, in turn, allows the detection of evolutionarily conserved functional associations (Gabaldón and Huynen, 2004). On the other hand, paralogs may provide valuable insights on the way in which genomes evolve and new genes acquire novel roles (see Chapter 2).

1.2.2 Methods to identify orthologs and paralogs

Since orthology and paralogy are terms directly derived from the evolutionary study of sequences, a phylogenetic analysis is, *a priori*, the best approach for its identification. However the (automatic) interpretation of phylogenies presents both methodological and conceptual obstacles in practice.

One of the most common procedures to predict orthology is the tree reconciliation approach (Page and Charleston, 1997), which is based on the comparison of a gene's tree with a known species phylogeny. In brief, tree reconciliation is performed by means of a parsimony algorithm that infers a new tree (reconciled tree) in which, considering the scenario with the minimum number of gene losses, makes the gene's tree topology fully compatible with the species tree.

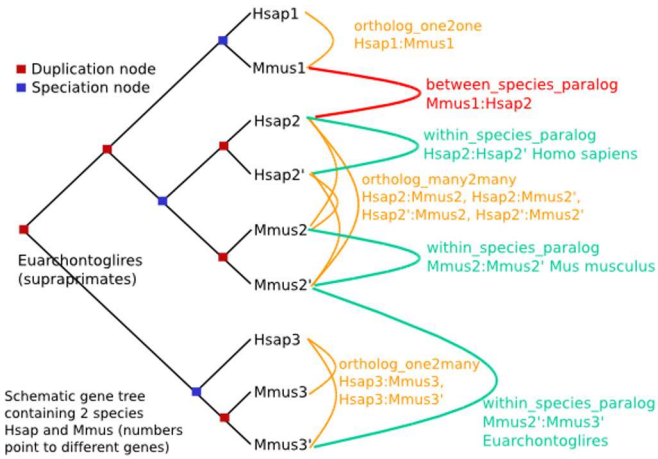


Figure 1.2: Example of a gene tree containing different orthology and paralogy levels. Figure was taken from the web site of Ensembl (Birney *et al.*, 2006)at http://ensembl.org.

This methodology entails at least three major drawbacks: First, the reconciliation algorithm relies on a completely resolved species tree, which is not always available (or certain). Secondly, even when the species phylogeny is well known, the expected relationships can be licitly violated by effect of the Lateral Gene Transfer process (LGT), which is quite frequent on procaryotes. Given that LGT can occur at any level of the species phylogeny, strict reconciliation algorithms tend to interpret this process as a big set of gene losses, resulting inwrongly predictedorthology relationships. In addition, complex patterns of gene gains and losses may also lead to similar results. Lastly, and similarly to the LGT problem, small artifacts on the gene phylogenies, very common in large phylogenies, usually lead to drastic changes on the reconciled tree. Recent developments have been implemented to allow the tree reconciliation algorithm to deal with some level of uncertainty in both the gene-tree and the species-tree. Among theses methods, we find soft-parsimony (Berglund-Sonnhammer *et al.*, 2006), model-based approaches (Arvestad *et al.*, 2003), and others using branch lengths and bootstraps to support tree partitions (Dufayard *et al.*, 2005). However, even with such improvements, the main drawbacks of tree reconciliation can not be totally avoided. For this reason, and because of the time cost associated to the application of phylogenetic methods, most genomic studies to date have preferred the use of alternative approaches such as the similarity based approaches.

Best Reciprocal (or bidirectional) Hits (BRH) (Huynen and Bork, 1998)is a popular methodology based on the assumption that a sequence from a given genome is more similar to its ortholog within another species genome than to any other sequence from such species. The reciprocity resides in the fact that if S1 and S2 are orthologs from two different genomes (G1 and G2), then S2 is expected to be the most similar sequence of S1 in genome G2 and S1 the most similar sequence of S2 in genome G1 . Computationally, this can be easily tested by the statistical assessment of all cross similarities between two genome sequences.

Problems associated to the BRH method are mostly derived from the violation of the assumption of similarity, which generally lead to a lack of accuracy (false negatives) or a deficit of coverage (since the method is limited to the detection of one-to-one relationships). The effect of LGT is, again, a main source of wrong orthology predictions. Many strategies have been developed to meet such deficiencies. For instance, the method known as Cluster of Orthologues Groups of proteins (COG) (Tatusov *et al.*, 1997), InParanoid (O'Brien *et al.*, 2005)and OrthoMCL (Li *et al.*, 2003) have extended the idea of symmetric best hits to more than two genomes and to more than one-to-one orthologs. Nevertheless, although these methods perform reasonably well in most cases, they have been shown to present many drawbacks that may lead to annotation errors or misinterpretation of data (Koonin, 2005; Eisen, 1998). In this thesis, an alternative phylogenomic strategy is explored to extend and improve the detection of evolutionary events from the phylogenetic tree interpretation (See 5.4)

1.3 Brief introduction to the phylogenetic pipeline

Independently from the scale of study, any phylogenetic analysis comprises several steps that range from the search for (putative) homologous sequences, the reconstruction of a multiple sequence alignment, and the inference of a phylogenetic tree. Next, a brief description of these basic phylogenetic steps is given.

1.3.1 Searching for sequences

As discussed in the previous section, sequence similarity is currently the more extended method to discover groups of (potentially) evolutionary related genes or proteins. This is, indeed, the first necessary step for any phylogenetic analysis. Although similarity can be regarded in many ways, most of the current approaches base their searches on the so called Local Alignment Algorithms. In brief, these algorithms attempt to find the regions between two sequences that maximize the number of identical, or chemically similar, residues. Regions similarity is then statistically assessed to obtain a confidence based on the size of the database use to perform the search. Such a value expresses the probability of finding by chance an alignment of the same level of sequence similarity using the same database. FASTA (Pearson and Lipman, 1988) and BLAST (Altschul *et al.*, 1990) are, by far, the most extended programs that implement this methodology.

1.3.2 Multiple Sequence Alignments

Multiple sequence alignments (MSA) are groups of sequences organized in the form of text matrices in which every sequence's residue is placed over its homologs in the remaining sequences. Currently, many programs and algorithms do exist that reconstruct MSA of hundreds or thousands of sequences(Edgar, 2004; Notredame *et al.*, 2000; Eddy, 1995). The quality of the alignments produced by such algorithms is generally good, however it is not rare to that different

programs render different alignments for the same sequence set. Recently, this has generated an important debate on how small variations in MSAs can lead to different topologies in the phylogeny (Wong *et al.*, 2008; Rokas, 2008). Besides the mutations affecting specific nucleotides, differences among sequences are very often caused by insertions and deletions of sequence stretches (indels). MSAs deal with indels through the insertion of "gaps" between certain residues. In practice, a gap within a sequence is interpreted as that a given position is missing in such sequence, but present in any other. How gaps are placed in a MSA represents the main source of differences among algorithms. Although the importance of indels is widely recognized, MSAs containing large regions of gaps are usually not very well modeled by the phylogenetic inference. In fact, programs usually treat gaps in an arbitrary way or just as non-informative positions. Consequently, an extended procedure is to include a cleaning step previous to the phylogenetic tree inference. Thus, portions of the alignment with a high content of gaps are removed, keeping only the best aligned parts. Several tools (Talavera and Castresana, 2007) have been developed to perform this task in an automatic way.

1.3.3 Phylogenetic tree inference

There are many different ways to infer a phylogenetic tree, and the range of possibilities covers from simple distance methods to complex Bayesian or Maximum Likelihood (ML) approaches. For years, the choice among different methods has been based on feasibility reasons. Thus, more accurate approaches such as ML and Bayesian methods have been commonly prevented. In more recent times, thanks to the availability of faster computer and the development of smarter algorithms, computationally demanding methods begin to be feasible and slowly tend to become standard in phylogenetic analyses.

There are three major approaches for phylogenetic estimation, namely distance methods, parsimony and statistical approaches (including Maximum Likelihood and Bayesian inference).

1.3.3.1 Parsimony methods

Maximum Parsimony (MP) is a classic method to infer phylogenies in which the preferred tree is the one that implies the minimal number of character changes along its branches. Characters can be any attribute that varies among the different elements included in the analysis. Such attributes can be morphological, physiological or even behavioral, but in the era of molecular phylogenetics, characters usually correspond to the nucleotides or amino acids that compose a biological sequence. In an exhaustive MP phylogenetic approach, all possible tree topologies are evaluated in terms of the number of character changes that are needed to explain the data, and the one with least changes is chosen as the preferred phylogeny. Computing all possible character change scenarios over a large number of sequences is computationally expensive and, since the number

of trees to evaluate grows, exponentially, with the number of sequences included, MP approaches over very large data sets are not usually feasible. In addition, besides its high computational demands, MP suffers from other important drawbacks such as the difficulty of dealing with multiple substitutions or homoplasy. For all these reasons the use of MP is not common in large scale analyses.

1.3.3.2 Distance methods

Distance methods, in contrast, are among the most popular tree construction methods. They are, by far, the fastest approach to build phylogenies and provide reasonable accuracies in terms of topology. Therefore they have long been the method of choice to reconstruct phylogenies of large groups of taxa or to conduct bootstrap calculations. The main disadvantage of distance methods, however, is the possibility of getting stuck in poor local minima. To reconstruct a phylogeny using a distance-based method, evolutionary distances between all pairs of sequences involved are computed and used to to approximate the evolutionary distance by calculating the percentage of non-identical sites between each possible pair of sequences. This approach is fast but it usually under-estimates distances between distantly-related sequences, due to the fact that multiple substitutions might have occurred at the same site. Therefore, distances are usually estimated from amino acid substitution matrices and then corrected to account for multiple substitutions. Once a distance matrix is obtained, several approaches can be followed to actually reconstruct the phylogenetic tree. Clustering methods such as UPGMA (Unweighted pair group method using arithmetic averages) or Minimum evolution , were the first to be designed. But the Neighbor joining (NJ) algorithm, is nowadays the most extensively used. This algorithm constructs the tree by clustering neighboring sequences in a stepwise manner. At each step, multiple topologies are examined and the one with minimal branch lengths is chosen. For large data sets it is only able to examine a small proportion of the total number of possible topologies. NJ is known to be statistically consistent; therefore if correct pair-wise distances are used, it reconstructs the true tree. Usually however, distance estimation is prone to statistical errors, compromising the accuracy of the resulting tree.

1.3.3.3 Statistical methods

The last major group, which comprises statistical methods, evaluates the appropriateness of the different trees by using a specific statistical framework. In brief, given an evolutionary model, they calculate the likelihood (or probability) that a given tree would have produced the observed sequence alignment. The evolutionary model is a set of probabilities for residue substitutions, differences in rates of evolution and any other parameter that can be used to describe how a given set of sequences may have evolved. The main difference between Maximum Likelihood (ML) and Bayesian Inference (BI) methods is their specific statistical framework. While ML computes the likelihood that a given tree would have produced the alignment, in BI the posterior probabilities of the trees, conditional to

that alignment, are considered by using Bayes theorem. Since both approaches are NP-hard problems, performing exhaustive searches on the topology space is not possible for moderate sizes of data and different heuristics are applied. Most ML implementations use hill-climbing algorithms to search the tree space. Changes in the topology are introduced at each step to subsequently evaluate the likelihood of the tree. The algorithm stops when a maximum likelihood tree is found. Significant progress in ML computation has been achieved with the release of fast and accurate programs such as PhyML Guindon and Gascuel (2003) and RAxML (Stamatakis *et al.*, 2005). At the side of Bayesian analysis, the program MrBayes is perhaps the most popular. It implements a Markov Chain Monte Carlo method for taking samples from the probability distribution of the phylogenetic tree space.

Chapter 2

Phylogenomics and Genome Evolution

2.1 Phylogenomics

Since the release of the first complete genomic sequence (Sanger *et al.*, 1977), and specially over the past decade, the number of sequences deposited in databases has grown exponentially. This has been largely favored by the development of new high throughput techniques such as the 'shotgun' sequencing method (Messing *et al.*, 1981). In fact, from the first non viral sequenced genome, that from *Haemophilus influenzae* (Adams and Fleischmann, 1995), and headed by the human genome project started in 1990 (Consortium, 2001), the number of fully sequenced genomes has almost doubled every year . According to the Genomes On-Line Database (GOLD) (Liolios *et al.*, 2006), 833 complete genomes are currently published, and 2924 sequencing projects are in progress.

Consequently, a transition toward a genomic perspective is taking place in many areas of biology. Phylogenomics, the field where this thesis is framed, is the name given to the area of bioinformatics that focuses on the phylogenetic study at a genomic scale. Initially, phylogenomic surveys were applied to the prediction of gene function (Eisen, 1998; Eisen *et al.*, 1997), but its use has been indeed extended to many areas.

2.2 Phylomes: scaling up the phylogenetic pipeline

A number of genome wide experimental and computational analyses have been performed to date that capture different aspects of the biology of the human cell. These analyses include, among many others, those of the so-called transcriptome (Suzuki and Sugano, 2006), proteome (Humphery-Smith, 2004), interactome (Gandhi *et al.*, 2006) and metabolome (Nielsen and Oliver, 2005). The

availability of such large data sets have added new dimensions to the study of organisms; not only are they useful in elucidating the function of otherwise uncharacterized proteins, but they also provide information on the system-level properties of the cell (Benner, 2003). The reconstruction of the evolutionary histories of all genes encoded in a genome is known as the species' phylome, and constitutes another source of genome-wide information.

The term "phylome" was first used in 2001, when Sicheritz-Pontén and Andersson reconstructed more than 8000 phylogenetic trees from seven microbial genomes using phylogenetic distance methods (Sicheritz-Pontén and Andersson, 2001). More recently, other studies have applied the same approach to the detection of co-evolution patterns (Gabaldón and Huynen, 2005). However, reaching a genome-wide scale on the phylogenetic analysis of complex eukaryotes still presents many computational and methodological challenges, at least if high quality results are aimed. , As a consequence, most current phylogenomic studies require finding a compromise between the sophistication of the phylogenetic pipeline, and the time needed to complete it. Besides obvious parameters such as the number and size of the sequences included in the analysis, other factors do affect the performance of the global analysis. The most important is probably the choice of the phylogenetic approach used to infer the trees. Although statistical methods (Bayesian or Maximum Likelihood) provide the best results, they become highly expensive in terms of memory and CPU time when they are applied at genomic scales. In fact, they can only be scaled up through the parallel computation of all the required analyses. Furthermore, it should be taken into account that any extra parameter, such as the evolutionary model and variation of rates among sites, implies doubling the number of required tests and, hence, the time of the analysis.

One of the goals of this thesis is to contribute in overcoming such challenges and finding novel approaches to efficiently reconstruct complete phylomes. As a reference, the reconstruction of the human phylome presented in chapter 5 took more than 715 million hours (23 years) of single CPU time (Xeon 2,4Ghz), which correspond to 60 days on 140 parallel computers.

2.3 Genome Evolution

Genomes encode the information that makes species and individuals different to each other. Studying the evolution of genomes provides us not only with a general view on the species relationships, but also with insights on how organisms gain new features or adapt to precise environments. Among other approaches, genome evolution can be addressed by combining the evolutionary information present in all the genes they encode. In this respect, the use of phylomes is particularly suitable.

In the second part of this thesis (Results), topics such as the phylogenetic variability among genes, the incidence of the gene duplication process, or the gene expression evolutionary dynamics after the duplication of a gene , are addressed at a genomic scale through the analysis of phylomes.

2.4 Phylogenetic variability at genomic level

Phylogenies, by definition, are expected to resolve the evolutionary relationships among a group of organisms. However, the data used to recreate such relationships may come from many sources, and, therefore lead into different results. Generally, the accuracy of the predictions increases with the number of homologous characters considered, given that a global consensus is expected among the majority of such characters. For this reason, molecular sequences, containing abundant and primary information about organisms, have been adopted as the best source to infer taxonomic relationships. Particularly, certain parts of the cell's DNA, such as the mitochondrial genome or the ribosomal genes, have received a special attention. Perhaps one of the best examples of this is that of the 16S ribosomal RNA sequence, which has been extensively used for taxonomic studies due to its ubiquitousness and high degree of evolutionary conservation. Nevertheless, even genetic data may produce ambiguous results. For instance, it has been shown that molecular phylogenies derived from the analysis of different gene families may lead to incompatible evolutionary conclusions about the organisms they belong to.

Even using the most powerful statistical framework to date, artifacts can not be totally avoided in a phylogenetic pipeline. As a consequence, a mandatory question in the genomic era is whether genome-wide phylogenetic analyses may help to reduce incongruence. Although one might think that artifacts and ambiguity would get diluted at a genome scale, an increasing number of analyses suggests that this is not the case. This topic, addressed in two recent reviews(Rokas and Carroll, 2006; Jeffroy *et al.*, 2006), plays a central role in chapter 5 of this thesis, whose results are also reviewed and framed in the context of accross-gene topological diversity in (Castresana, 2007).

2.5 Gene duplication: a key process on the evolution of genomes

Gene duplication plays a central role in the evolution of genomes, and, consequently, in the evolution of species. As originally proposed by Ohno(Ohno, 1970), this process is one of the most important sources for the acquisition of novel gene functions, and its study provides insights on the way in which genomes diverge.

The duplication of a gene implies that its original function becomes initially redundant. In principle, if no direct benefit exist from two copies coding for the same function, evolutionary forces are expected to suppress such redundancy by the pseudogenization and eventual elimination of one of the duplicates. However, before pseudogenization takes place, mutations and random drift on one of the duplicated genes may lead to the acquisition of new functional features or to the sub-division of the ancestral one between the duplicates. If this occurs, both copies become indispensable and, therefore, their retention is favored. Different hypotheses exist to explain such process, most of them being not mutually ex-

cluding. For instance, neo-functionalization models imply the acquisition by one of the duplicates of a novel function, which was not present in the common ancestor. In contrast, sub-functionalization models (Lynch and Force, 2000) assume that each duplicate will retain only part of the functionality of the ancestor. In any case, the result of both processes would be the same: there will be a selective pressure to retain both duplicated copies.

2.6 Duplicates and the evolution of gene expression

Not only mutations affecting coding regions are responsible for functional changes. Those affecting the regulatory elements of the gene constitute an additional and very important source of functional variation. Gene regulation involves the control of when, where and how a gene is expressed. This is the reason why gene expression has been proposed to play a determinant role on the development and differentiation of species since the early times (King and Wilson, 1975). Moreover, genomic surveys have shown that changes in gene expression may cause phenotipic variations . In this respect, the evolutionary view of gene expression may provide insights on the global understanding of gene regulation. For this, variation of gene expression profiles can be studied in the context of gene duplication by applying the same rules used for the neo- and sub-functionalization models. If regulation of gene expression is subjected to the same forces that affect the coding sequences, as proposed by earlier studies , sub- and neo-functionalization models (Force *et al.*, 1999) would predict that duplicated genes will tend to derive towards complementary or new expression patterns that prevent them from being deleted. A suitable scenario for such type of analysis is that of spatial or temporal variation in the expression pattern of duplicates, given that in such cases a change in gene role is especially evident. Chapter 7 of this thesis presents a genomic study in which results and methods from chapter 5and 6 are combined to analyze several evolutionary aspects of the variation of human and mouse tissue expression patterns.

Chapter 3

Thesis overview

Throughout the following four chapters of this thesis, a set of phylogenomic surveys are presented that covers from the reconstruction of the human phylome to its analysis from different perspectives.

Chapter 5 describes the reconstruction of a high quality version of the human phylome, including 39 eukaryotic organisms from the plants and fungi to humans. More than 20,000 phylogenetic trees are inferred for each of the 5 evolutionary models tested. Best fitting model trees are used to assess several uncertain branch points in the eukaryotic tree of life, to discard cases of lateral gene transfer, and to derive a extensive catalog of orthologous genes.

In Chapter 6, gene duplication process is quantitatively addressed across nine important periods along the evolution of the 39 eukaryotic organisms used in chapter 1. To this end, a novel method that estimates the age of duplications events based on the topological scanning of phylogenetic trees is evaluated by comparing predictions with synteny based results. Subsequently, the method is applied to the set of duplications derived from the human phylome to obtain the duplication frequency for each period. Results reveal a duplication rate profile that is compatible with the proposed existence of at least one round of Whole Genome Duplication proposed preceding the origin of vertebrates. In addition, duplicates from different periods are functionally characterized and compared.

Chapter 7 extends the study of the gene duplication process by exploring the evolution of spatial gene expression patters between human and mouse duplicates. The process is addressed from the perspective of the sub and neo-functionaliciation models, regarding spatial variations on the the gene expression patterns as a functional change. An especial attention is given to the temporal frame in which duplication, and neo or sub-functionalization occur. Results yield no correlation between duplicates functional specialization and the time of the duplication event.

Finally, in Chapter 8, two new phylogenomic resources are presented: a database of complete phylomes, phylomeDB; and a set of programming libraries (ETE). Both resources are intimately associated and provide a framework to

perform large-scale phylogeny-based analyses and to exploit large collections of precomputed data.

Chapter 4

Objectives

- Overcome the technical challenges associated to the reconstruction of high quality and complete phylomes through the compilation of a first version of the human phylome.
- Explore the possibilities of the analysis of complete phylomes to reduce the uncertainty on some species relationships.
- Develop methods for the automatic interpretation of large collection of phylogenies, and explore its possibilities for the automatic prediction of orthology and paralogy relationships.
- Provide insights on the global understanding of the gene expression evolution between duplicated genes
- Provide a public phylogenomic resource that allow the research community to exploit the use of phylomes.

Part II

Results

Chapter 5

The Human Phylome

The complete set of gene phylogenies associated to an organism is known as the organism's phylome Sicheritz-Pontén and Andersson (2001). Phylomes provide us with the evolutionary history of all the genes encoded in a genome, and therefore, with the relationships among their homologs in other species.

It is known that the quality of a phylogenetic analysis is highly dependent on the method used to identify (potentially) homologous sequences, align them, and infer their a relationships tree. However, given that the most accurate approaches are very time-consuming, the use of sophisticated pipelines at genomic scale has been traditionally prevented by a large computational cost. For this reason, current efforts have focused on a limited set of model species and less accurate methodologies. Most remarkably projects are the Ensembl database Birney *et al.* (2006), now including phylogeny based orthology predictions; the TreeFam Li *et al.* (2006) and HOVERGEN Duret *et al.* (1994) databases, providing automatically derived and curated phylogenies of animal gene families; and the Adaptive Evolution Database (TAED) Roth *et al.* (2005), focusing on phylogenetic detection of adaptive events in gene families. Although all of them provide a valuable source of evolutionary information, the still miss several important features which should be present in a high-quality phylogenetic analysis: e.g. alignment cleaning steps, evolutionary model testing, statistical branch support or gene-based perspective.

In the present chapter, we address the reconstruction of the human phylome by applying the most accurate phylogenetic methods currently available, aiming to compile the first high-quality catalog of all human gene evolutionary histories. In contrast to previous attempts, we use a gene-based approach that aims at maximizing both the coverage over the human genome and the taxon-sampling among fully sequenced eukaryotic genomes. As a result, building the human phylome presented here took two months on a total of 140 64-bit processors, which is roughly equivalent to 23 years in a single processor. To our knowledge, this represents the most sophisticated phylome reconstruction pipeline and the largest computing time investment for a single phylome reported to date.

5.1 Phylome scope and phylogenetic pipeline

The human phylome presented here is derived from the proteins encoded by 39 publicly available eukaryotic genomes (Table 5.1). This set is particularly rich in metazoan species (19 species, 50%), including 14 chordates, 3 arthropods and 2 nematodes. The second largest group is that of fungi, comprising 11 species and thus making a total of 30 opisthokons. The remaining group includes eight species from diverse *phyla*, among which are one amoebozoan (*Dictyostellum discoideum*), two plants (*Arabiopsis thaliana* and *Chlamydomonas reinhardti*), two apicomplexans (*Plasmodium falciparum* and *Plasmodium briggsae*), and three excavates (the diplomonad *Guillardia theta* and the kinetoplastids *Leishmania major* and *Paramecium tetraurelia*). This distribution of species makes our set especially suitable for addressing the evolution of protein families among the opisthokonts. It covers, therefore, a period that is rich in important evolutionary innovations, from the origin of apoptotic pathways Blackstone and Green (1999)to the emergence of complex communication patterns Fisher and Marcus (2006).

To derive a phylome from the above-mentioned proteome database we applied a phylogenetic pipeline to each human protein. This fully automated pipeline (described in more detail in the Materials and methods section. Chapter 10) emulates the manual workflow used by phylogeneticists: from sequence, through alignment, to phylogenetic reconstruction. It starts with a sequence search against the proteome database to retrieve groups of significantly similar proteins that are then aligned. Multiple sequence alignments are automatically trimmed to remove gap-rich regions and used for phylogenetic inference. Such a phylogenetic reconstruction combines NJ, ML and Bayesian methods. Firstly, a NJ tree is constructed with BioNJ Gascuel (1997), and secondly, this NJ tree is used as a seed in a ML analysis using PhyML Guindon and Gascuel (2003). In the ML analysis, up to five different evolutionary models were tested for each tree using a discrete gamma-distribution model with four rate categories plus invariant positions. Both the gamma shape parameter and the fraction of invariant positions were estimated from the data. Finally, the ML tree rendered by the best fitting model, as determined by the Akaike Information Criterion (AIC) Akaike (1973), was further refined with a Bayesian approach as implemented in MrBayes Ronquist and Huelsenbeck (2003). After the Bayesian analysis, a consensus tree was produced by using the 'halfcompat' option of MrBayes, which produces a topology in which all partitions are compatible with at least 50% of the trees produced by the Monte Carlo Markov Chain analysis. Genome-wide application of the pipeline resulted in 21,588 final alignments and 129,510 trees covering different phylogenetic models and approaches.

5.2 Evolutionary model selection

Both ML and Bayesian analyses are model-based approaches that can provide divergent results when different evolutionary models are assumed. Several authors

Code	Species name	source	Proteins included	trees
Hsa	Homo sapiens	Ensembl	21,726 (99.1%)	21,588 (100.0%)
Ptr	Pan troglodytes	Ensembl	17,113 (79.3%)	19,577 (90.7%)
Mmu	Macaca mulatta	Ensembl	19,285 (89.2%)	19,765 (91.6%)
Mms	Mus musculus	Ensembl	19,934 (78.9%)	18,825 (87.2%)
Rno	Rattus norvegicus	Ensembl	18,675 (85.7%)	18,585 (86.1%)
Cfa	Canis familiaris	Ensembl	16,657 (91.8%)	18,834 (87.2%)
Bta	Bos taurus	Ensembl	18,457 (79.9%)	18,736 (86.8%)
Mdo	Monodelphis domestica	Ensembl	17,004 (80.7%)	18,013 (83.4%)
Gga	Gallus gallus	Ensembl	12,325 (66.5%)	15,758 (73.0%)
Xtr	Xenopus tropicalis	Ensembl	14,721 (60.6%)	15,787 (73.1%)
Tni	Tetraodon nigroviridis	Ensembl	14,896 (53.4%)	14,585 (67.6%)
Fru	Fugu rubripes	Ensembl	15,834 (72.3%)	15,155 (70.2%)
Dre	Danio rerio	Ensembl	16,042 (74.9%)	14,808 (68.6%)
Cin	Ciona intestinalis	Ensembl	5,588 (50.9%)	9,421 (43.6%)
Aga	Anopheles gambiae	Ensembl	6,131 (43.0%)	9,310 (43.1%)
Dme	Drosophila melanogaster	Ensembl	6,812 (49.6%)	9,771 (45.3%)
Ame	Apis mellifera	Ensembl	4,484 (33.4%)	8,616 (39.9%)
Cel	Caenorhabditis elegans	Ensembl	5,826 (29.8%)	8,190 (37.9%)
Cbr	Caenorhabditis briggsae	Integr8	5,171 (39.2%)	7,899 (36.6%)
Ago	Ashbya gossypii	Integr8	2,020 (42.8%)	3,603 (16.7%)
Cal	Candida albicans	Other	2,733 (33.8%)	3,899 (18.1%)
Cgl	Candida glabrata	Integr8	2,129 (41.1%)	3,627 (16.8%)
Cne	Cryptococcus neoformans	Integr8	2,532 (38.5%)	4,102 (19.0%)
Dha	Debaromyces hansenii	Integr8	2,302 (36.5%)	3,885 (18.0%)
Ecu	Encephalitozoon cuniculi	Integr8	626 (32.8%)	1,203 (5.6%)
Gze	Giberella zeae	Integr8	3,076 (26.4%)	4,412 (20.4%)
Kla	Kluyveromyces lactis	Integr8	2,077 (39.1%)	3,715 (17.2%)
Ncr	Neurospora crassa	Other	2,521 (23.7%)	4,221 (19.6%)
Sce	Saccharomyces cerevisiae	Ensembl	2,317 (35.1%)	3,769 (17.5%)
Spb	Schizosaccharomyces pombe	Integr8	2,421 (48.8%)	4,102 (19.0%)
Yli	Yarrowia lipolytica	Integr8	2,487 (38.1%)	4,152 (19.2%)
Ddi	Dictyostelium discoideum	Integr8	3,843 (29.4%)	5,165 (23.9%)
Ath	Arabidopsis thaliana	Integr8	9,450 (26.6%)	5,390 (25.0%)
Cre	Chlamydomonas reinhardtii	Other	2,303 (11.7%)	3,504 (16.2%)
Gth	Gillardia theta	Integr8	161 (35.7%)	458 (2.1%)
Pfa	Plasmodium falciparum	Integr8	1,330 (25.3%)	2,507 (11.6%)
Pyo	Plasmodium yoelii	Integr8	1,188 (15.3%)	2,272 (10.5%)
Lma	Leishmania major	Integr8	2,082 (26.0%)	3,130 (14.5%)
Pte	Paramecium tetraurelia	Integr8	140 (30.2%)	345 (1.6%)

Table 5.1: For each species: the 'Proteins included' column indicates the number of proteins present in trees of the human phylome and the percentage they represent; and the 'Trees' column indicates the number of trees in the phylome with proteins from that species (and the percentage from the phylome it represents). 'Source' indicates the database from which the protein data for that species were retrieved

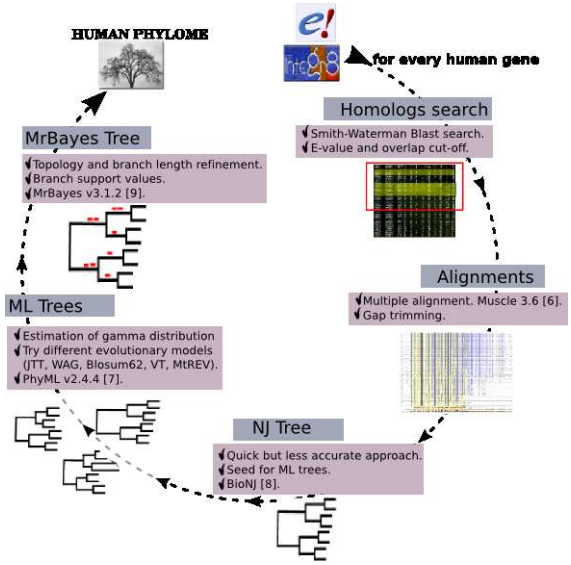


Figure 5.1: Schematic representation of the phylogenetic pipeline used to reconstruct the human phylome. Each protein sequence encoded in the human genome is compared against a database of proteins from 39 fully sequenced eukaryotic genomes to select putative homologous proteins. Groups of homologous sequences are aligned and subsequently trimmed to remove gap-rich regions. The refined alignment is used to build a NJ tree, which is then used as a seed tree to perform a ML likelihood analysis as implemented in PhyML, using four different evolutionary models (five in the case of mitochondrial encoded proteins). The ML tree with the maximum likelihood is further refined with a Bayesian analysis using MrBayes. Finally, different algorithms are used to search for specific topologies in the phylome or to define orthology and paralogy relationships.

have shown that the use of an appropriate model is crucial for the reconstruction of correct phylogenies and that the origin of the sequences involved (that is, the range of organisms involved) is not always a good predictor of the most appropriate model Keane *et al.* (2006); Bruno and Halpern (1999). Applying a wrong evolutionary model to a given data set might even lead to the reconstruction of wrong phylogenies with a high support Buckley and Cunningham (2002).

To avoid such pitfalls, we tested using the ML approach several models that are complementary in their scope, namely: JTTJones *et al.* (1992), a general model for globular, nuclear-encoded proteins; BLOSUM62 Henikoff and Henikoff (1992), inferred from protein blocks of 62% sequence identity; WAG, derived from a database of globular proteins with a broad range of evolutionary distances Whelan and Goldman (2001); and VT, based on amino acid replacement rates suited for distantly related sequences Müller and Vingron (2000). Additionally, phylogenies of the proteins encoded in the mitochondrial genome were also reconstructed using mtREV, a model that has been specifically designed for this kind of data Adachi and Hasegawa (1996). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used. The gamma parameter and the fraction of invariant positions were estimated from the data.

Among the models tested, JTT was chosen as the best fitting model in a majority of the trees (14,683, 68.0%), followed by WAG (6,388, 29.6%), Blosum62 (461, 2.1%) and VT (26, 0.1%). MtREV was chosen as the best model in ten out of the thirteen mitochondrial-encoded human proteins. Surprisingly, the phylogenies of subunit 6 of NADH dehydrogenase and subunits 1 and 2 of cytochrome oxidase were best fitted by JTT, Blosum62 and WAG models, respectively. To assess whether a tree produced by the NJ approach has sufficient predictive value for the model selection step, we compared the model chosen by the full ML approach (that is, reconstructing a ML phylogeny for every model) to the model selected when the likelihood of the seed NJ tree was assessed under different models, allowing for branch-length optimization. In 86.7% of the cases the model chosen by both methods was the same.

This confirms and extends earlier results Keane *et al.* (2006) and, more importantly, suggests that the pipeline could be simplified by basing the model selection on the tree produced by BioNJ.

5.3 Topological diversity within the human phylome

Recent advances in resolving the tree of eukaryotes are converging into a model that comprises a few large super-groups Keeling *et al.* (2005). Despite the general agreement on the classification of these major groups, several relationships, both among and within the different groups, remain controversial. In recent years, a number of large-scale approaches have been developed that combine the information obtained from several genes to resolve evolutionary relationships. Among these, the construction of super-trees and trees based on concatenated alignments

are among the most widely used Delsuc *et al.* (2005). These trees are useful in that they constitute a straightforward way of visualizing the combined phylogenetic signal of genes that are widespread in the species considered. However, it has been claimed that these trees are representative of only a small fraction of the genes encoded in a given genome, and that gene-sampling effects might lead to biased results supporting a specific species phylogeny Jeffroy *et al.* (2006); Dagan and Martin (2006).

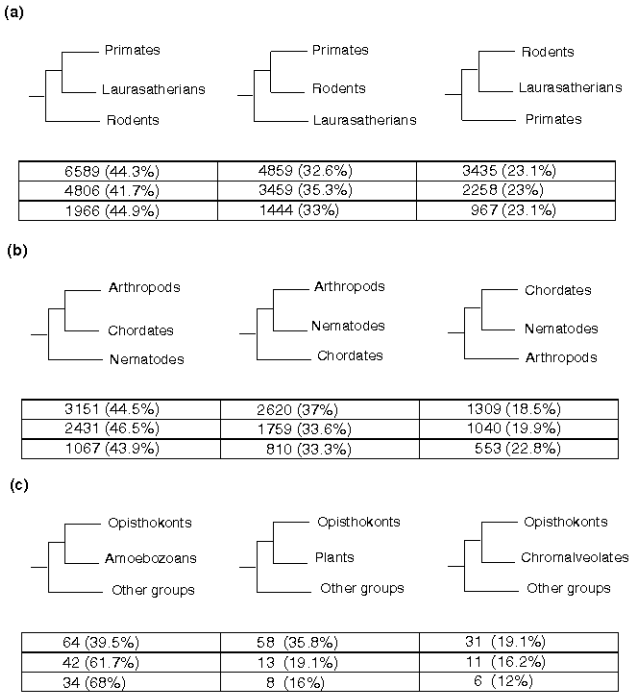
A phylome represents a broader, yet more complex to interpret, reconstruction of the evolution of an organism, since it comprises the phylogenies of all its genes. Most notably, the availability of a phylome opens the possibility for studying the relationships among species in a different way: that of quantifying the fraction of individual phylogenies whose topologies are consistent with a given hypothesis. Here we explored this methodology by specifically contrasting a number of evolutionary relationships that are controversial to some extent. We chose three different scenarios for which there is some level of controversy in the literature and that involve three different depths of the eukaryotic tree (Figure 5). Namely, the relative positions of nematodes, chordates and arthropods, the relationships among rodents, primates and laurasatherians, and, lastly, the grouping of opisthokonts with amoebozoans. To scan for phylogenies compatible with the different hypotheses, we adapted a previously described algorithm Gabaldón and Huynen (2003).

5.3.1 Ecdysozoa versus coelomata hypotheses

Perhaps one of the most debated issues regarding the tree of eukaryotes is the relative position of arthropods, nematodes and chordates. Traditionally, comparative anatomy placed arthropods and chordates in the coelomata clade, which contained animals with a true body cavity, while pseudocoelomates such as nematodes occupied a more basal position. However, phylogenetic analyses of 18S and 28S rRNAs supported an alternative view that grouped nematodes and arthropods, dubbed ecdysozoa, to the exclusion of chordates Aguinaldo *et al.* (1997). Since then, numerous multi-gene phylogenetic studies that support either of the hypotheses have been published (see, among others, Philippe *et al.* (2004); Dopazo and Dopazo (2005)).

Our results (Figure 5.2b) show a preponderance of genes whose phylogeny is consistent with the Coelomata hypothesis. Of the 7,080 phylogenies in the human phylome with representatives from the three groups, 3,151 (44.5%) support the Coelomata hypothesis, placing nematodes at a basal position, compared to 2,620 (37%) and 1,309 (18.5%) that group nematodes with arthropods (Ecdysozoa hypothesis) or with chordates, respectively. The relative fraction of trees supporting each topology is similar if we consider only the 5,230 trees with the highest topology support (posterior probabilities higher than 0.9 in the nodes grouping the considered taxa (Figure 5.2b).

Since the algorithm treats each gene individually, a certain level of redundancy exists because protein families with many members in the human genome



6589 (44.3%)	4859 (32.6%)	3435 (23.1%)
4806 (41.7%)	3459 (35.3%)	2258 (23%)
1966 (44.9%)	1444 (33%)	967 (23.1%)

3151 (44.5%)	2620 (37%)	1309 (18.5%)
2431 (46.5%)	1759 (33.6%)	1040 (19.9%)
1067 (43.9%)	810 (33.3%)	553 (22.8%)

64 (39.5%)	58 (35.8%)	31 (19.1%)
42 (61.7%)	13 (19.1%)	11 (16.2%)
34 (68%)	8 (16%)	6 (12%)

Figure 5.2: The alternative phylogenetic relationships among the taxa involved in the three evolutionary hypotheses considered. (a) Placental mammals: primates, laurasatheria and rodents. (b) Ecdysozoa versus Coelomata hypothesis: relationships among arthropods, chordates and nematodes. And (c) the Unikont hypothesis: relationship among opisthokonts, amoebozoans and other eukaryotic groups. The numbers indicate the number of trees supporting each topology. For each alternative topology numbers on the top row refer to the total number of trees with a given topology, and what percentage of the total it represents; numbers in the middle row refer to those trees for which the posterior probabilities of the two partitions shown in the figure are 0.9 or higher. Numbers in the bottom row refer to the number and percentage of gene families supporting each topology.

contribute more trees to the phylome. These would affect the topological analysis if there are great differences in the distribution of family sizes supporting each topology. To correct for this redundancy we grouped the individual gene-trees into families if their seed sequences appeared together in a tree. Then each family was considered to support a single topology. If more than a single topology was supported, the one supported by a majority of members was chosen. As shown in Figure 5.2 (bottom row), the percentage of families supporting each topology is similar to the results obtained when genes are treated individually.

The finding that all three possible topologies, including the one widely considered as wrong in the literature, are supported by a significant number of trees, illustrates the inherent difficulty of resolving the species phylogeny from gene phylogenies. We have found similar topological diversity in the three scenarios considered (see below) and also, to smaller degrees, in apparently undisputed evolutionary relationships (results not shown).

Similar results showing variability in the relative positions of arthropods, nematodes and chordates have also been found in topological analyses of the phylogenies of 507 eukaryotic orthologous groups Wolf *et al.* (2004) and of 100 protein families Blair and Hedges (2005). These deviances from the species phylogeny might be the result of different processes, including convergent evolution or varying evolutionary rates. In the case of the Ecdysozoa and Coelomata hypotheses, the accelerated rate of evolution in the nematode sequences has been proposed as the main cause preventing the acceptance of the Ecdysozoa hypothesis. For instance, some studies have shown that when fast evolving genes are removed from the data set, the ecdysozoa group is accepted with high confidence Philippe *et al.* (2004); Dopazo and Dopazo (2005). Therefore, the relative abundance of the different topologies should be considered with caution, since differences in evolutionary rates, if they are widespread, could result in a majority of the gene trees supporting a wrong species phylogeny.

5.3.2 Relationships among placental mammals

The phylogenetic relationship among placental mammals has attracted great interest in recent years Murphy *et al.* (2004). A still open question is the relative grouping and branching order of the groups rodentia, primates, lagomorpha, artyodactyla and carnivora. Four of these groups are represented in the present phylome, namely primates (human, chimpanzee and macaque), artyodactyla (cow), carnivora (dog) and rodents (rat and mouse). While the monophily of artyodactyla and carnivores, both belonging to laurasatheria, is largely undisputed, the crucial question is whether rodents have a basal position relative to the other groups or whether they join primates on a common node. analyses of concatenated alignments from nuclear genes are consistent with the rodents being a basal group and primates being monophyletic with laurasatheria Kullberg *et al.* (2006); Misawa and Janke (2003). However, phylogenies based on mitochondrial genes as well as the common presence of several mutational events and the insertion of MLTA0 elements support the clustering of primates and rodents to the exclusion of laurasatheria Murphy *et al.* (2004); Thomas *et al.* (2003).

In our analyses the results seem to favor the basal position of rodents, although the difference with the alternative hypothesis of a clade grouping rodents and primates is not great (Figure 5.2a). From the 14,883 trees in the human phylome with representatives for the three groups (Figure 5.2a), 6,589 (44.3%) show a topology in which rodents are basal, compared to 4,859 (32.6%) and 3,435 (23.1%) trees in which rodents are monophyletic with primates and laurasatheria, respectively. As in the case of arthropods, nematodes and chordates, all possible topologies are fairly represented. Here too, differences in the relative evolutionary rates, and the possible long-branch attraction effect, might have an effect on the high proportion of trees showing rodents at a basal position, since rodent sequences have been shown to have the highest rates of substitutions when compared with primates and artiodactyls Ohta (1995); Zhang (2000).

5.3.3 Unikont hypothesis

Among the most difficult problems in the evolution of eukaryotes is resolving the relative branching order of the major eukaryotic groups. The evolutionary distances and the level of sequence divergence involved results in a star-like tree with the major eukaryotic groups branching out in a poorly defined order. Nevertheless, phylogenetic analyses have been used to cluster some of the groups. One such case is the union of amoebozoans and opisthokonts, dubbed the unikonts Cavalier-Smith (2002). Evidence supporting this group comes from phylogenies based on concatenated alignments of up to 149 genes Philippe *et al.* (2004) as well as from morphological data. However, this grouping is still not widely accepted among systematicists. In the present analysis a single amoebozoan genome, that of *Dictyostellum discoideum*, has been included, together with representatives from three other major groups, including excavates (*L. major*, *P. tetraurelia*, *G. thetha*), plants (*A. thaliana*, *C. reinardthii*) and chromoalveolates (*P. falciparum*, *P. yoelii*).

We scanned the phylome for trees supporting the grouping of opisthokonts with each of the other major groups, provided that at least four of the five major groups were represented in the tree (Figure 5.2c). Of the 165 trees in the human phylome including at least four of the five major groups, 64 (39.5%) supported the Unikont hypothesis. The alternative hypotheses of opisthokots being monophyletic with either plants, chromoalveolates or excavates are supported by 58 (35.8%), 31 (19.1%) and 9 (5.6%) trees, respectively. However, differences between the Unikont and the alternative hypotheses are greater when only the 68 trees with high (>0.9) posterior probability in the partition supporting the monophyly are considered. In this case the Unikont hypothesis is consistent, with 42 (61.7%) trees compared to 13, 11 and 2 trees supporting the alternative hypotheses of opisthokonts grouping with plants, chromoalveolates and excavates, respectively.

5.4 Phylogeny based orthology detection.

Although an increasing number of genome-wide experimental data sets is becoming available for human, most experimental analyses are performed in model species such as mouse, fruit fly, yeast and the nematode *Caenorhabditis elegans*. Additionally, for historical or practical reasons, alternative model species are used to investigate specific systems or pathways. Such is the case with the use of *Neurospora crassa*, *Yarrowia lipolytica* and *Bos taurus* models in the characterization of the multiprotein enzyme NADH:Ubiquinone oxidoreductase (Complex I), in which an intricate evolution and the use of different naming schemes in the various species complicate the transfer of knowledge among investigators studying the different model species Gabaldón *et al.* (2005). Comparative genomics can be used for transferring functional information across species, a process that requires the establishment of evolutionary relationships among genes encoded in the different genomes. Such relationships are best established by means of detecting orthology, rather than just homology. Orthologs are a special case of homologous genes that diverged from a common ancestor through speciation events, in contrast to paralogs, which originate from duplication events Fitch (1970).

The large amount of phylogenetic information present in a species phylome can be exploited to establish the orthology relationships of all its genes. The fact that phylogeny-based algorithms perform the analysis over whole phylogentic trees allows to extend prediction to all the species included in the phylome (some of them, as in the presented human phylome, not model-species). In addition, phylogeny based methods might augment the resolution of current predictions by taken into account multi-gene relationships. The most extended procedure to extract the evolutionary events present in a phylogeny is tree reconciliation, which consists on the comparison of the gene tree with a fixed species tree. However, considering the high degree of topological diversity observed in the human phylome, we reasoned that any algorithm based on reconciliation would inevitably infer a false duplication events in the trees showing topologies that depart from the canonical species tree. Therefore, we decided to explore an alternative, fully automated approach that does not require a fully resolved species phylogeny and a reconciliation phase. Our method (described in 11.1) uses the level of species overlap between two connected branches to decide whether their parental node represents a duplication or speciation event, and hence it allows for a greater level of congruence. We applied the algorithm over the best fitting model gene phylogeny, identifying a total of 33,787 orthology relationships (data can be downloaded from `http://phylomedb.bioinfo.cipf.es/`).

In order to evaluate the quality of our phylome-based predictions, we performed a comparison against several alternative methods by using a recent reference data set comprising 67 human-mouse and 45 human-worm orthologous pairs from five multi-gene families Hulsen *et al.* (2006). Considering the size of the families and the intricate evolutionary histories involved, this reference set should be considered a highly stringent test. For each of the methods compared we computed the sensitivity, which is a measure of the coverage over the reference set, and the positive predictive value, which is the proportion of correct orthology

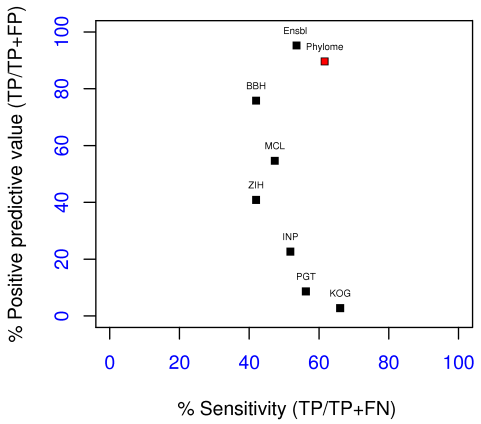


Figure 5.3: Benchmarking comparison of different orthology inference algorithms. The reference set used in the benchmark of Hulsen et al. Hulsen *et al.* (2006) is taken as a gold standard to compute the number of true positives (TP), false positives (FP) and false negatives (FN) yielded by each method. For each method the sensitivity (S = TP/(TP+FN)) and the positive predictive value (P = TP/(TP + FP)) are computed. Methods described in Hulsen *et al.* (2006) are indicated as BBH (Best reciprocal hits), MCL (OrthoMCL), ZIH (Z-score 1-hundred.), INP (Inparanoid), PGT (phylogeny-based algorithm used in van Noort *et al.* (2003)), KOG (Clusters of eukaryotic orthologous groups). 'Phylome' represents the results of our pipeline and algorithm, and Ensbl the orthology relationships predicted by Ensembl database

predictions, that is, the number of true positives over the sum of true positives and false negatives. The results of the benchmark showed narrow differences in terms of sensitivity (Figure 5.3). All methods are able to predict only about half (40% to 66%) of the orthologous pairs in the reference set. Our method scores second best, with 61.6% sensitivity compared to 66.1% for the clusters of eukaryotic orthologous genes (KOG) method; Ensembl reaches a coverage of 55.57%. As we noted before, this low coverage reflects the inherent difficulty of the reference set, in which manual orthology assignments have taken into account domain organization analysis and other sources of information. Most remarkable are the big differences encountered in the positive predictive values. These range from 2.8% (KOG) to 86.61% (our algorithm) and 95.24% (Ensembl). Altogether, the results show that phylogeny-based orthology detection methods can provide substantial improvement in terms of positive predictive value when sophisticated phylogenetic pipelines are implemented.

5.5 Absence of horizontal transfers of eukaryotic genes in the human lineage

The extent and scope of horizontal gene transfer (HGT) events among organisms has been the subject of intense debate Kurland *et al.* (2003). The emerging view is that HGT constitutes an important process of evolution in prokaryotes and that it is more restricted, if not virtually absent, in eukaryotes. However, as more eukaryotic genomes are being sequenced, the number of putative cases of gene transfers in eukaryotes is growing. Reported cases include acquisition of prokaryotic genes Andersson *et al.* (2003); Ricard *et al.* (2006); Goldsmith *et al.* (2005) and transfers of mitochondrial genes between plants Bergthorsson *et al.* (2003) and between animals Alvarez *et al.* (2006). Horizontal gene transfer in the human genome has been addressed in the past. For instance, after the initial sequencing of the human genome the claim was made that up to 223 bacterial genes, likely acquired by HGT, could be found in the human genome Consortium (2001). This claim, however, was later rejected on the basis of phylogenetic analysis Salzberg *et al.* (2001). The existence of horizontally transferred genes from other eukaryotes in the human genome has never been reported despite the fact that integrative viral sequences can migrate between vertebrate species and that these viruses can sometimes carry genes within their sequences, making the hypothesis theoretically plausible Bromham (2002).

The species represented in our phylome include organisms that are tightly linked to human, either because they are pathogens (plasmodium and several fungi), or used as a source of food (cow, yeast). A recent transfer from any of these species to the human genome could, in principle, be detected as a human protein being placed in a 'wrong' phylogenetic context. However, caution must be taken when interpreting phylogenies, since such topologies can also be explained by alternative processes such as multiple gene-loss or lack of phylogenetic resolution.

To find such putative cases we scanned the human phylome to detect trees in which the phylogenetic position of the human seed protein could suggest a possible HGT event. For this purpose we applied a series of increasingly stringent filters. These filters (Table 5.2) consisted in identifying trees in which: the human seed protein has non-primate proteins as nearest phylogenetic neighbors; such topology cannot be explained simply by the loss of the orthologous sequences in the other primates or multiple losses in mammalian groups; the partition suggesting the HGT is supported by a high posterior probability (>0.9) in the Bayesian analysis; and that partition is also supported by ML analysis. This methodology bears some similarity to that proposed by Hallet *et al.* (2004) Hallet *et al.* (2004) in that it specifically defines possible scenarios for HGT.

A total of 99 trees (0.47%) passed the first two filters, thus having a topology that could be explained by an HGT event. However, only 8 of these trees had a posterior probability supporting the HGT partition of 0.9 or higher in

Filter	Filter description	Number of cases
1	Non primate protein as nearest neighbor	
2	Topology cannot be explained by gene loss	99
3	Partition is supported by a posterior probability >0.9	8
4	Partition is supported by both ML and Bayesian analysis	0

Table 5.2: Filters used to detect potential cases of HGT and the number of trees meeting such filters.

the Bayesian analysis, and none of these was supported by the ML analyses, indicating that the partitions suggesting the horizontal transfer are not strongly supported. We interpret these results as a lack of evidence supporting the existence of human genes originating from recent horizontal transfers from the lineages considered and argue that the observed HGT-like topologies are rather the result of phylogenetic artifacts. This interpretation is consistent with the generally adopted view that horizontal gene transfers among multi-cellular eukaryotes is virtually absent due to the existing natural barriers that prevent transferred genes from reaching the germ-line Kurland (2005).

Chapter 6

Dating duplicates

Gene duplication is one of the most important mechanisms by which genomes acquire novel functions Ohno (1970). Not only have recent genomics surveys provided evidence for the abundance of duplicated genes in all organisms Vogel and Chothia (2006), but also it has been observed that gene duplication is often associated with processes of neo- and sub-functionalization Roth *et al.* (2007). Thus, during the course of evolution, gene families can increase their size and functional scope through gene duplication events Tatusov *et al.* (1997). However, while certain periods in the evolution of genomes have shown to be clearly associated to duplications involving the entire genome (whole genome duplications), others show different degrees of independent gene duplications. The time associated to a gene duplication event is usually estimated by means of the synonymous substitution rate (dS) among the resulting paralogs. We argue, however, that dS can only be used as reliable age estimation for relatively short periods of time, beyond which the signal gets saturated.

In the present chapter, we explore the incidence of the gene duplication process across different periods in the evolution of the human genome by using an alternative dating method based on the analysis of gene phylogenies. Moreover, we attempt to characterize the main functional roles in which duplicates associated to each period are involved.

6.1 dS as an estimator of divergence time

The degree of sequence divergence between two related protein-coding genes can be expressed as the amount of substitutions occurred at the level of their DNA sequences. These substitutions can be subdivided into non-synonymous or synonymous depending on whether the nucleotide change results in a different amino acid in the resulting protein. Since synonymous substitutions are considered free of selective pressures, at least at the protein level, these are very often assumed to linearly increase with time. Accordingly, synonymous substitution rates between

two sequences (dS) are extensively used as a proxy for their divergence time. In the particular case of two paralogous genes, emerged by duplication, dS ratio is generally considered a good estimator for the time of the duplication event. However, although many recent surveys use this approach to study the evolution of duplicated genes Makova and Li (2003); Gu *et al.* (2002); Li *et al.* (2005), the central assumption that dS ratios rise linearly with time is never tested for the ranges considered. Besides possible differences in synonymous substitution rates of biological ground , many methodological issues, such as the saturation of the signal and alignment uncertainty, may result in large differences in dS estimates between pairs of genes duplicated at the same time.

In the light of the mentioned caveats, alternative methods such as duplication dating based on the topology of phylogenetic trees may prove more adequate. In the next section, we evaluate a modified version of the species-overlap algorithm, described in chapter 5 and section 11.1, used to estimate the evolutionary time in which a gene duplication occurs. Such method, based on the species content of each tree branch, is described in more detail in 11.4.

6.1.1 Topological age estimation versus synonymous substitution rates

Deciphering the time in which an ancestral gene got duplicated is a difficult task that can only be addressed by indirect analyses. In this respect, knowledge on a whole gene duplication event (WGD) may provide a good calibrating point to assess the accuracy of the methods that are usually applied to estimate the age of a single gene duplication, since genes originated from a WGD should be all equally mapped to same time. In order to explore the drawbacks attributed to the use of dS as a proxy to the divergence time, and to explore an alternative phylogenetic based method, we have performed a variability analysis based on the WGD occurred in the saccharomycotina phyla about 100Ma ago.

The set of paralogous gene pairs resulting from such WGD was extracted by a previous study based on the syntenic analysis of several yeast genomes Byrne and Wolfe (2005). Conservation of gene order among orthologs of the species appeared after the WGD is a quite reliable sign that such genes originated from the same event. Moreover, in order to keep only the most evident cases, we limited our analyses to the 250 paralogous pairs in which both copies were conserved in the three post-WGD species (*Sacaromices cerevisiae*, *Sacaromices castellii* and *Candida glabrata*) (see 9.2).

Parallely to the dS estimation (see method 10.3), we used a recent yeast phylome, reconstructed in Marcet-Houben and Gabaldón (2007), to perform a phylogeny-based age estimation of the 250 paralogs. First, we applied the species-overlap algorithm to detect the same paralogy relationships in the yeast phylome. Since each duplication event represents a given node in the gene phylogeny, the topological age was established according to the sequences (organisms) grouped by such node. In our case, three major groups were considered:

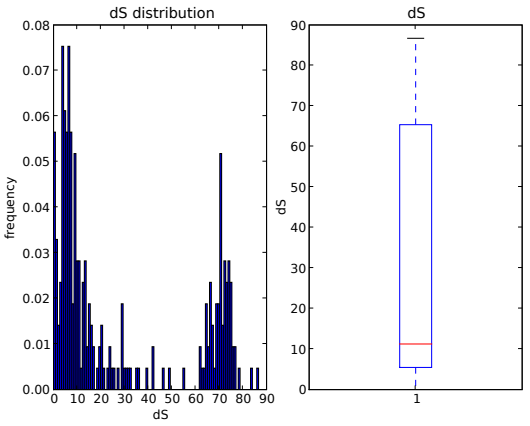


Figure 6.1: Distribution of dS values obtained for 250 yeast paralogs mapped to the same whole genome duplication event. dS was calculated by maximum likelihood as implemented in the the codeml program Yang (1997).

1. High-Evidence-WGD, if the duplication involves sequences from the three saccharomycotina species and no other organism is present.

2. Post-WGD, if the duplication involves only sequences from the saccharomycotina species but not all of them are represented.

3. Pre-WGD, if the duplication involves sequences from an organism prior to the saccharomycotina group.

The distribution of dS ratios obtained for the 250 paralogous pairs is shown in Figure 6.1. Surprisingly, the range of values covers from 0.004 to 86.71, with a mean of 27.94 and a high standard deviation (28.61). Such a variability exemplifies the degree of uncertainty associated with this measure.

On the other hand, phylogeny-based predictions yielded 110 duplications (~50%) exactly mapped to the High-Evidence-WGD group. 23 pairs (~10%) were assigned to earlier lineages (post-WGD) because of the gene loss in any of the organisms. A significant number of paralogs (30,8%) were assigned to the tree branch just before the speciation of of the WGD species (pre-WGD). Finally a total of 20 pairs (9%) were mapped to farther branches (pre-WGD). Figure 6.2 shows the distribution of such numbers across the fungal tree of life. Interestingly, a similar variability of dS ratios can be observed for the paralogs mapped to different branches in the tree (corner plots in figure 6.2).

Overall, our results suggest that, given the high level of uncertainty observed, the general assumption of linearity between dS ratios and divergence time should be taken carefully. In fact, the bi-modality observed in the dS distribution evidences a elevated rate of saturated predictions (60<dS<=90). By contrast, the phylogenetic approach tested here seem to yield more adjusted and interpretable predictions. Although only the 50% of the paralogs were exactly assigned to

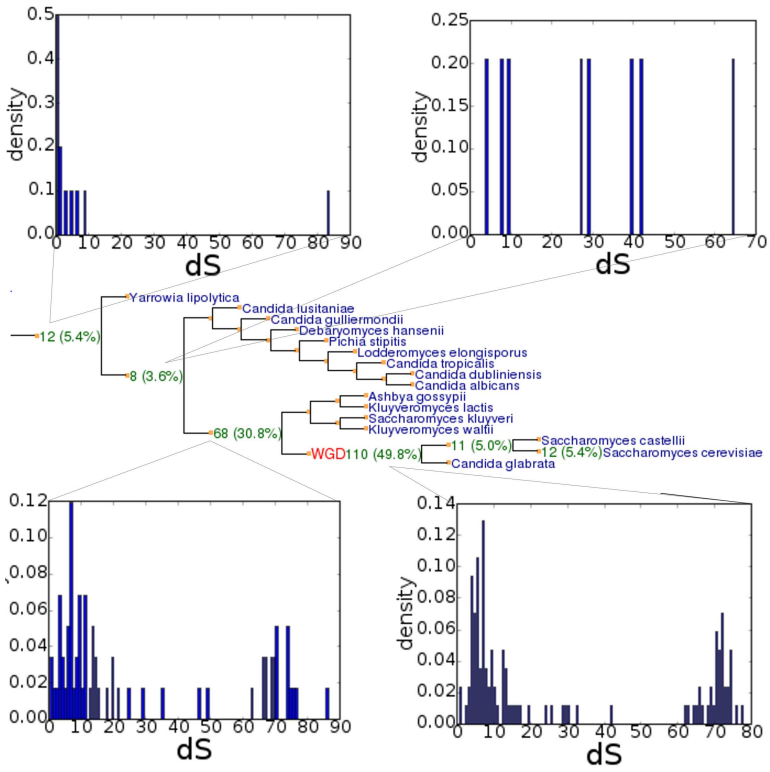


Figure 6.2: Comparison among the topological age prediction of 250 Yeast paralogs and their distribution of dS values. Number in green are the amount of paralogs pairs were assigned to each branch by a phylogenetic analysis. The 4 plots in the corners are the normalized histograms for the dS values associated to the paralogs mapped to each branch.

the same time pointed by the synteny analysis, the dispersion of values seems to be much lower than in the case of dS. For instance, the 90% of predictions concentrates at only 1 branch step from the correct evolutionary event in the tree. Therefore, we consider this method to be a good alternative to previous approaches.

6.1.2 dS variability among human duplicates.

In the same line of the previous analysis, we wondered on the variability of dS rations among all human duplicates. To this end, we used the human phylome to extract all duplication events leading to the human linage. For each duplication, dS was calculated among the resulting paralogs. Similarly to the results obtained for yeast, the range of values obtained was surprisingly high (figure 6.4).

On the other hand, we estimated the phylogenetic age of all human paralogs by considering nine phyla groups that represent the major branching points in

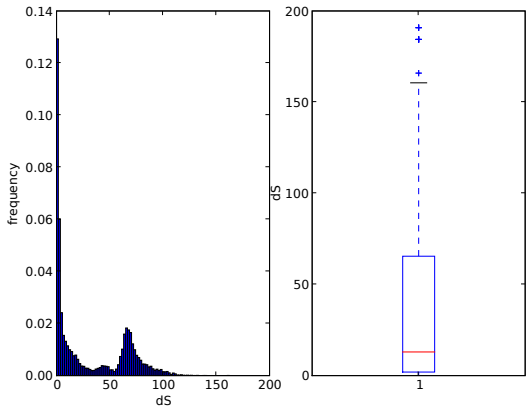


Figure 6.3: Distribution of dS values among the human paralogs detected in the human phylome.

the linage leading to human. These nine evolutionary periods range from the origin of Opisthokonts to the separation of human and primates lineages.

Figure 6.4 shows dS ratios distribution for all human duplicates pairs arranged according to their topological age. It can be noticed how big dS overlaps are found between different categories. The problem is especially evident for duplications occurred before the expansion of mammals, were the dS ratio signal seems to be completely saturated. However, large overlaps can be found even within the most recent evolutionary periods, where dS is not saturated. Thus similar dS values can be found for duplicates topologically mapped both to the origin of human linage and to the origin of primates.

Overall, these results show the caveats of using dS rates as a proxy for estimating divergence time. Although measures based on topological dating such as the one proposed by us are far from perfect, they appear to efficiently map evolutionary events to discrete periods of time. It must be noted, however, that a correct topological mapping relies on the quality of the under laying phylogenetic tree, the absence of horizontal transfer events, the broadness of taxonomic sampling and our level of understanding of the phylogeny of the species involved.

6.2 Lineage-specific gene duplication in the Human Phylome

To quantify the extent of gene duplication that has occurred in the lineages leading to human, we applied the species-overlap described in 11.1 to identify all duplication events present in the human phylome. Each event was associated to one of nine major branch points by following the topological age estimation method (see 11.4).

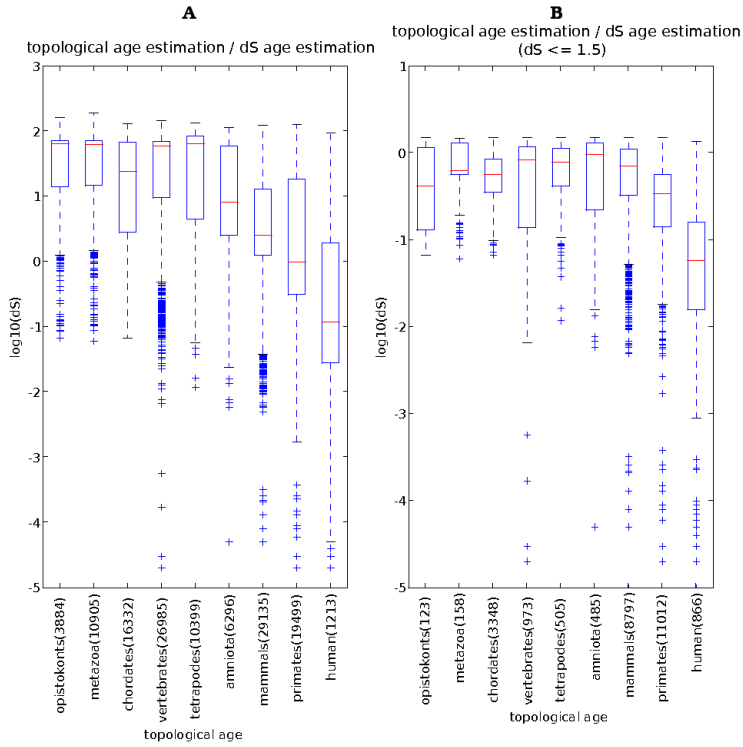


Figure 6.4: Distribution of dS ratios between duplicated gene sequences, arranged according to nine evolutionary periods. dS is expressed as the log10(dS). Each paralogous gene pair was assigned to one of nine evolutionary categories by the topological analysis of its phylogeny (see 11.4).

Redundant cases, derived from the analysis of gene phylogenies from the same family, were avoided by calculating the rate of duplications per gene. Such duplication frequency was estimated as the quotient between the total number of duplication events mapped to a given stage and the number of trees rooted at a deeper branching point. This result in a number that be interpreted as the number of duplication events occurred per gene at each one of the evolutionary periods. For example, if 100 duplication events are detected at the chordates level, and 50 phylogenetic trees exist that are rooted to older sequences (metazoa, fungi or basal eukaryote sequences), the rate of duplications per gene at chordates level would be 2.0. Results obtained for the analysis of the human phylome are shown in Figure 6.5B.

The highest peak in gene duplication events corresponds to the base of chordate evolution, after the split of urochordates (Ciona intestinalis) and vertebrates. This observation is consistent with previous results supporting the existence of at least one round, and probably two rounds, of whole genome duplications before the radiation of vertebrates Panopoulou *et al.* (2003); Blomme *et al.* (2006), which could explain the increase in phenotypic complexity of vertebrates rela-

tive to other chordates such as cephalochordates (amphioxus) and urochordates (Ciona).

The second largest peak appears at the base of the metazoans, after their split with fungi. The relatively large duplication rate (0.58 duplications per tree) at this point could be interpreted as a result of a WGD at the base of metazoan evolution or, alternatively, an accumulation of smaller scale duplications. To the best of our knowledge, the possibility of a WGD at the base of metazoan evolution has not been proposed in the literature Meyer (2003) and we believe it deserves some deeper consideration in future analyses. If the WGD scenario is considered, then an extensive gene loss should have followed it, since the duplication rate here is lower than the one found at the base of vertebrate evolution. The alternative scenario would assume a high number of smaller scale duplications that affected more than 50% of the genes. These duplication events would have accumulated over the period of time extending from the split of fungi and metazoans to the split of chordates and other metazoans.

Also remarkable is the relatively high duplication rates found in the lineages leading to mammals, primates and hominids. This suggests that duplications have played a major role in the evolution of these groups, something that has already been noted from comparisons of primate genomes Bailey and Eichler (2006).

6.3 Functional trends among duplicated gene sets in the Human Phylome

It is described that gene duplication might result in the amplification and/or diversification of the biological processes in which duplicates are involved. Therefore, inspecting the functions of gene families that have undergone duplication at different evolutionary stages may provide clues about the processes that played roles in the major transitions that occurred during those stages.

To detect such functional trends we searched for Gene Ontology (GO) terms that are significantly over-represented in the set of genes that underwent duplications during the different stages of eukaryotic evolution. We performed this analysis automatically with the aid of the program Fatigo+, from the Babelomics suite Al-Shahrour *et al.* (2006). At each evolutionary stage (Figure 6.5A) we compared the annotations of the duplicated human genes with those of the rest of the human genome. We selected those terms whose over-representation was significant based on a false discovery rate test (adjusted p-value < 0.00001).

The present analysis detects over-represented functions within the 'Biological Process' GO category among genes duplicated at different evolutionary periods (Figure 6.5C). It is, therefore, different from complementary analyses that detect functional shifts and different patterns of amino acid replacement among duplicated pairs Abhiman and Sonnhammer (2005); Seoighe *et al.* (2003). Interestingly, these complementary analyses also show differences among functional classes. In most evolutionary stages, we found several terms from different GO

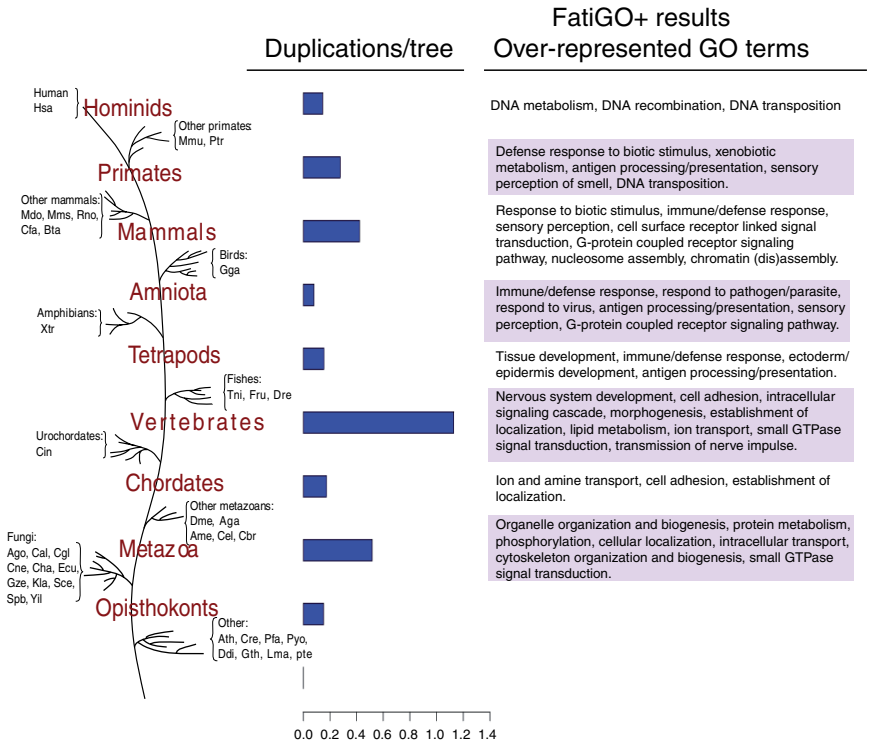


Figure 6.5: Estimates for the number of duplication events occurred at each major transition in the evolution of the eukaryotes. Species abbreviations are the same as in table 5.1. Horizontal bars indicate the average number of duplications per gene. Boxes on the right list some of the GO terms of the biological process category that are significantly over-represented compared to the rest of the genome in the set of gene families duplicated at a certain stage.

levels and categories that are significantly over-represented. Of these, some are specific to a given evolutionary transition (for example, lipid metabolism in vertebrates), while others are over-represented in a series of consecutive stages (for example, small GTPase signaling cascade).

Providing links between the over-represented terms and the functional or morphological transitions characteristic of each stage is not straightforward. Nevertheless, some terms do suggest the expansion of some physiological processes at a given evolutionary time. Terms related to maintenance of complex cellular structures, such as 'organelle organization and biogenesis', 'cytoskeleton', 'cellular organization' or 'cellular localization', are over-represented in genes duplicated before the divergence of fungi and metazoans, suggesting major transitions in cellular organization common to all opisthokonts.

The expansion of the process 'small GTPase signal transduction' in almost all major stages from the origin of opisthokonts to the vertebrates indicates a continuous expansion of signaling cascades that is likely related to the increasing level of multi-cellularity and tissue differentiation observed at these evolutionary stages. Similarly, protein families related to 'G-protein coupled receptor signaling pathway' were expanded before the amniota and mammalian radiations.

Also remarkable are the consecutive waves of expansion observed for the 'immune response' and related terms. They have occurred at every split from the origin of tetrapods to the origin of primates and suggest an increasing sophistication of the immune system. Xenobiotic metabolism terms are also over-represented in genes duplicated in primates. As noted before Bailey and Eichler (2006), the sophistication of the immune response and xenobiotic recognition and detoxification might have facilitated adaptation to changes in food sources and infectious agents.

The specific association of terms such as 'transmission of nerve pulse' or 'nervous system development' with families duplicated just before the vertebrate expansion is consistent with the development of a complex nervous system as compared to that of simpler chordates. Later on, the expansion of 'sensory perception' and related terms in the lineages leading to amniota, mammals and primates indicates increasing sophistication of the senses. Similarly, the term 'epidermis development' is over-represented in genes duplicated in tetrapods. This might be related to major skin modifications, which potentially allowed the conquering of the terrestrial environment by this group.

Chapter 7

Expression divergence among differently aged gene duplicates

Variations on the gene expression patterns can be regarded as functional changes. In particular, and although any kind of variations might be considered, changes in the spatial expression pattern denote a clear functional differentiation. Thus, the expression divergence between duplicated genes could be perfectly addressed through the neo- and sub-functionalization models proposed originally by Force (Lynch Force, 1999). Under either of such models, gene expression would be expected to play an important role on the retention of duplicates, since small changes on the original expression pattern may turn both duplicated genes indispensable for the cell.

Many recent studies have focused on testing some of theses predictions, specially regarding the relationship between family size, sequence divergence and expression breath (number of tissues in which a gene is expressed). Sub- and neo-functionalization models predict that, if the acquisition of complementary spatial patterns prevents duplicates from pseudogenization, a reduction of expression breadth is to be expected after each round of duplication. As a result, expression breadth is likely to be narrower in larger families. This fact was indeed observed by several recent studies Freilich *et al.* (2006); Huminiecki and Wolfe (2004) as a negative correlation between expression breath and family size. In addition, Gu *et al* Gu *et al.* (2004) showed that the expression diversity tends to be higher for duplicated genes than for single copy genes.

Another aspect that has been addressed, but for which some controversy remains, is that of the time frame in which gene duplication and subsequent expression divergence occur. For instance, in order to favor gene retention, the process of tissue specificity acquisition should follow the gene duplication shortly, preceding the eventual pseudogenization of one of the duplicates. In this direction,

some studies have addressed the relationship between the age of the duplication, usually estimated from the differences in neutral sites between duplicates (dS), and their specific patterns of tissue expression. Gu et al Gu *et al.* (2002)found a significant positive correlation between the level of coding sequence dissimilarity and expression divergence among yeast paralogs. Makova and Li Makova and Li (2003) analyzed tissue expression pattern divergence among human duplicates and reported a linear correlation between sequence divergence and spatial expression difference. However, other studies have reported a lack of significant correlation between sequence divergence and expression divergence. The differences in the previous studies may be explained by different methodologies to measure expression divergence and the lack of sufficient resolution of the use of synonymous substitutions to estimate the age of the duplication.

In this chapter, we exploit the availability of the human phylome to tackle the previous questions by applying full phylogeny-based method to detect and date duplication events. This approach allow us to account for the actual complexity of duplications and gene-loss scenarios by differentiating among one-to-one, one-to-many and many-to-many paralogy relationships. We also rely on the topological dating of the evolutionary events (see Chapter 6), which extends the range of ages and avoids the saturation effect derive from synonymous substitutions rates. Using these data, we explore the relative timing of a duplication event and that of the acquisition of specific tissue expression patterns. In particular we investigate whether different evolutionary periods are associated with different degrees of tissue expression divergence and found that genes duplicated at ancient periods may get specialized to tissues that originated only recently.

7.1 Tissue specificity, expression breadth and complementarity in human paralogous families

We scanned the complete set of 21,588 human gene phylogenies included in the human phylome and detected 7522 human duplication events. Tissue expression patterns were inferred from the expression data sets of 79 healthy human tissues retrieved from the SymAtlas project at GNF (see Material and Methods, 9.3).We represent an expression profile as an array of 0's (indicating the gene is not expressed) and 1's (indicating the gene is expressed) over the 79 tissues considered. The expression breadth of a given gene is defined as the number of tissues in which it is expressed. Genes with a narrow expression breadth are tissue specific. To account for one-to-many and many-to-many paralogy relationships, expression profiles of paralogous sets involving more than one gene were grouped into a single profile in which 1 indicates the expression of one or more genes of that set . The total number of tissues in which any of the genes of a particular paralogous set is expressed is defined as the global expression breadth.

We then explored the correlation between the number of paralogs involved in a duplication event (duplication size) and the expression breadth of such genes. In line with what has been reported by others Huminiecki and Wolfe (2004); Freilich

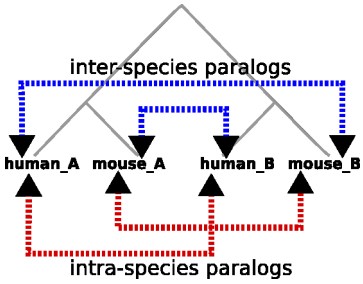


Figure 7.1: Example of inter- and intra-species paralogy relationships.

et al. (2006) we observe a weak but significant (Spearman R=0.16 p=6.23e-44) positive correlation between the global expression breadth and the number of paralogs. In contrast, a negative correlation (Spearman R=-0.18 p=3.6e-55) is found between the average expression breadth per gene and the number of paralogs. Both correlations suggest that, in general, the larger the number of paralogs resulting from gene duplications, the greater the number of tissues covered by their expression and the higher the tissue specificity of each of the components. In other words, larger families will be expressed in a higher number of tissues but each member, separately, is likely to be more tissue specific.

For each duplication event, we determined the complementarity level between the tissue expression profiles of each of the paralogous gene sets (C, see methods 11.5). In 5128 (67%) duplications, the degree of tissue expression complementarity (C) is greater than 0, meaning that the expression patterns of the duplicates differ in, at least, one tissue. From these cases, 838 (11%) show full expression complementarity (C=1), that is, the two sets of paralogs have not been observed to be expressed in the same tissue.

7.2 Differences in expression between orthologs and paralogs

To test whether the high levels of tissue expression divergence observed among paralogs is associated to the event of duplication and is not just a result of their divergence in time, we compared the levels of tissue expression asymmetry of paralogs to that of orthologs. For this, we exploited the availability of mouse tissue expression data in the SymAtlas set at GNF to define a reduced expression data set, including 29 homologous human-mouse tissues. Expression profiles were computed as described above. To avoid possible biases derived from comparing intra-species paralogies against inter-species orthology relationships, we only considered in this case human-mouse paralogies derived from the same duplication event (inter-species paralogies, Figure 7.1).

This resulted in 3502 human-mouse paralogs pairs (from duplications occurring at the origin of mammals or later) that were compared to 4426 human-mouse

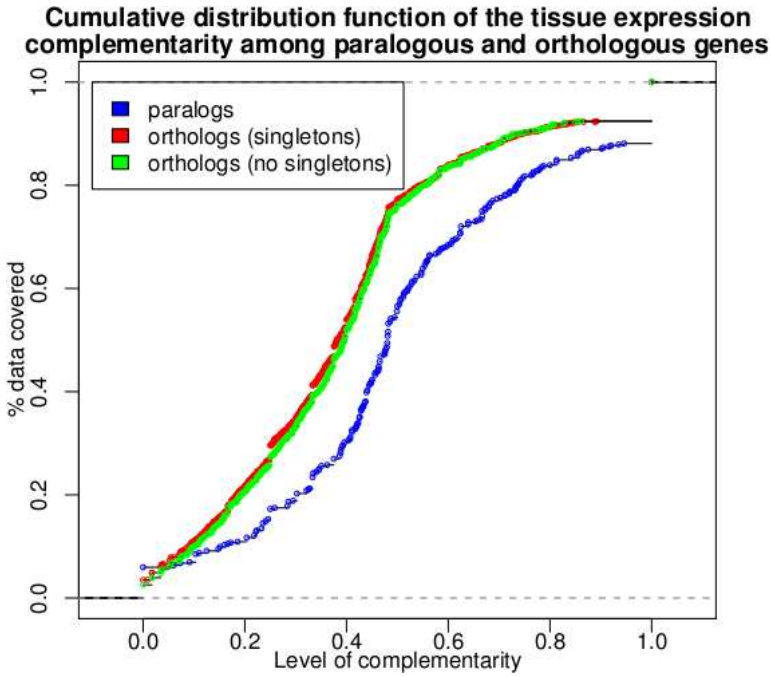


Figure 7.2: Cumulative distribution functions of the tissue expression complementarity of paralogs and orthologs. Distributions biased to the bottom-right corner show higher levels of complementarity. The blue line is the distribution shown by paralogous gene pairs. The red and green lines represent the distribution of orthologs singletons and non singletons respectively. Differences between paralogs and orthologs distribution are supported by a significant (pvalue=0.0, D=0.25) Kolmogorov-Smirnov test.

orthologs pairs in terms of their level of tissue expression complementarity. Our results (Figure 7.2) show that the distribution of complementarity levels between inter-species sets of paralogs is significantly higher (KS-test: pvalue=0.0, D=0.25) than that of a) singleton human-mouse orthologs, b) non-singleton human-mouse orthologs. Nevertheless, although lower as compared to that of paralogs, relatively high levels of complementarity can be also found between orthologous pairs of genes.

7.3 Gene duplication is directly followed by higher levels of tissue expression divergence

That paralogs have higher levels of tissue expression divergence relative to orthologs of similar age does not necessarily mean that this level of divergence was acquired shortly after the duplication. Alternative scenarios may involve slow but continuous accumulations in tissue expression patterns since the time of duplication. To test this we used the set of gene duplications occurred at the base of the mammalian lineage and that have several representatives in both human and

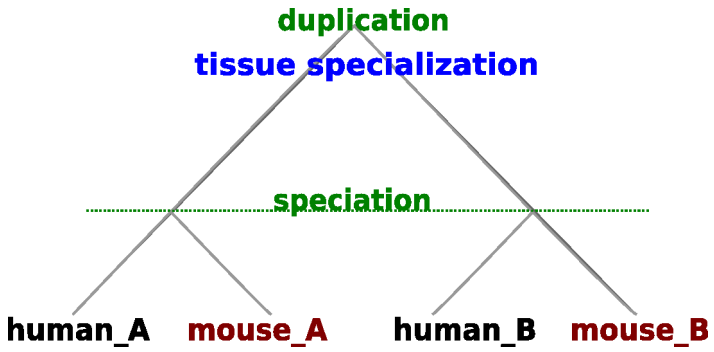


Figure 7.3: Schematic representation of the early acquisition of the tissue expression complementarity among human and mouse paralogs.

mouse genomes. In this case, the comparison of the tissue expression patterns between intra- and inter-species paralogs can provide information on whether the observed level of complementarity was gained before or after the speciation event. Indeed, if the complementarity was gained prior to speciation, one would expect that the complementarity observed between human-human paralogs parallels that observed between mouse-mouse paralogs that emerged from the same duplication. In the opposite scenario, that of acquisition of complementarity after the speciation event, one would expect independence between the complementarities that were acquired in the human and mouse lineages. Our results show a significant correlation between the complementarities found between all types of intra- and inter-specific paralogous sets compared , suggesting that the expression complementarity between duplicates was mostly acquired prior to the human-mouse speciation event , and tends to be conserved in evolution. Note that this finding is in agreement with the idea that gene retention was favored by the acquisition of divergent tissue expression patterns. Indeed, only a relatively rapid acquisition of tissue expression divergence would have rendered both duplicates indispensable before the eventual pseudogenization of one of the copies.

7.4 Ancient duplications and modern specificities

We next explored the relationship between the age of gene duplication events and their implication in tissue expression complementarity. Our intention was twofold: firstly, we tried to solve current discrepancies on whether tissue expression divergence increases with the age of the duplicates and, secondly, we wondered whether the appearance of multi-cellularity and tissue differentiation in the different evolutionary periods is coupled to the emergence of tissue-expression complementarity of genes duplicated at that period. We found differences in complementarity levels among the duplicates originated at different evolutionary periods . For instance, duplications at the level of chordates, vertebrates, mammals and tetrapods show the highest levels of complementarity between the resulting duplicates. However, we observed no clear linear correlation between the age of

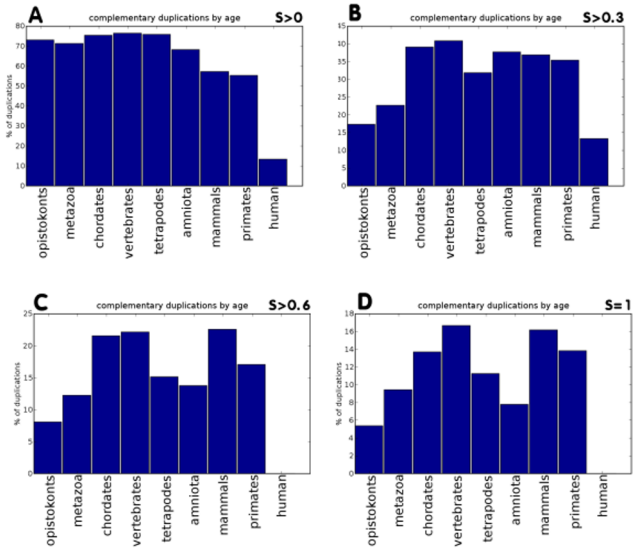


Figure 7.4: Tissue expression complementarity by duplication age. A, B, C and D represent 4 increasing degrees of complementary threshold. Each bar represent different evolutionary origins of the paralogs tested.

the duplicates and their degree of tissue expression divergence, that is, levels of complementarity do not increase or decrease with the age of the duplication .

Figure 7.4 shows the relative number of duplications associated to different degrees of tissue expression complementarity (C) for several evolutionary periods. Full complementarity (C=1) can be found between duplicates that date back to the common ancestor of opisthokonts, a unicellular organism in which no tissue differentiation could have occurred . Also remarkable is the complete absence of paralogs with a full complementarity (and even with C>0.3) in the most recent duplications in our set, those which occurred after the split of humans and chimpanzees.

Additional examples that suggest that duplication and tissue specialization are not necessarily coupled in time, include cases in which the full expression complementarity is linked to the differentiation of a single tissue, and the duplication is older than the tissue itself. As shown in Figure 7.5, many duplications involved in the differentiation of placenta, adaptive immune system, pancreas, liver and kidney, are more ancient than the origin of these tissues. Interestingly, a recent study have suggested that the development of placenta tissue was mostly mediated by ancient genes that were co-opted during its evolution, which would be compatible with our findings. Overall, these results suggest that gene duplication events and tissue expression divergence are not necessarily coincident in time. This would support the idea that tissue-expression differentiation is not necessarily coupled with the time of duplication. In other words, gene duplications are "not on demand". The retention of duplicate genes during the period that precedes the emergence of the corresponding tissues must thus be explained

by other types of sub-functionalization such as expression under different environmental conditions or function divergence. During periods in which tissues became differentiated, the whole pool of current duplicates would be susceptible to be involved in such kind of expression divergence. Nevertheless, duplications that occurred at evolutionary periods with higher tissue development would be more likely to be involved in complementarity before they get lost or get subspecialized by other factors. This could explain why duplications occurred at evolutionary periods in which more novel tissues and organs emerged (chordates to mammals) have higher proportions of duplications with full complementarity than duplications occurred at periods in which tissue differentiation was not prevalent (fungi, metazoa, primates, human)

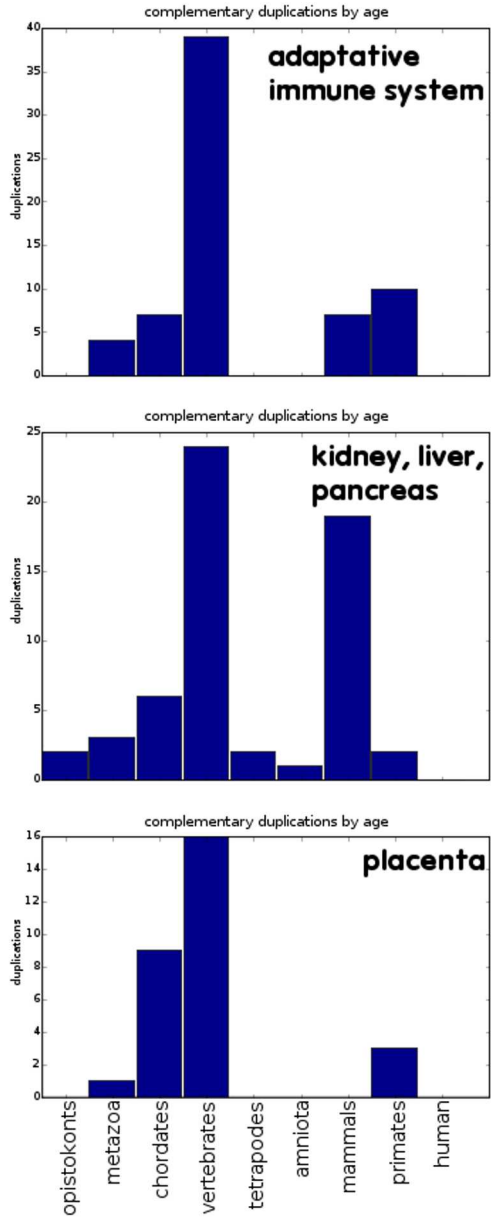


Figure 7.5: Duplication age profile for three groups of human paralogous genes which one of the duplicates shows expression specificity in the tissue(s) considered in each graph.

Chapter 8

PhylomeDB and the Environment for Tree Exploration

In chapter 5 we have shown the feasibility of reconstructing complete and high quality phylomes. We have also shown how its public availability is valuable for phylogenticists and molecular biologists. Phylomes can indeed be exploited either for large scale analysis or be used as encyclopedias of gene evolutionary histories. Although some databases provide automatically-derived and curated phylogenies , these follow a family-based approach, since they first group the genes into families and subsequently build a single phylogeny for each family. Moreover, the selection of species they include is determined by the specific scopes of each project. The automatic reconstruction of phylomes in a reasonable time span provide us with the possibility of adapting the species range and phylogenetic parameters to any specific scenario. Thus, a species phylome fitting custom needs can be reconstructed, although with large computational resources, in no more than few weeks. This constitutes a valuable source of information valid to address new biological questions. In this chapter, we present two phylogenomic resources: a public database for phylomes, phylomeDB; and the Environment for Tree Exploration (ETE), a programmable tool to visualize, manipulate and analyze phylogenetic trees.

8.1 PhylomeDB, a database of high quality gene phylogenies

PhylomeDB is a database that hosts genome-wide collections of single gene phylogenies testing a variety of organisms and covering different evolutionary scenarios. Multiple sequence alignments and phylogeny based prediction of orthologs are

also available to download. The Current version of PhylomeDB comprises 172,324 phylogenetic trees and 36,289 multiple sequence alignments covering species such as Human, Yeast, *E.coli* or Fly. It is accessible as a web portal in which database resources can be downloaded or interactively visualized. PhylomeDB content is arranged in phylomes, each of them presenting its own pipeline configuration and organisms scope. Thus, different resources belonging to different phylomes can be found for the same gene. Each phylome includes a description page which enumerates the organisms it covers and the proteome versions used to obtain protein or nucleotide sequences. It also details the peculiarities of its phylogenetic pipeline and the programs and parameters used in each step.

8.1.1 PhylomeDB structure

Phylomes hosted in phylomeDB are encoded following a common format for the species, sequence, and phylome identifiers. This allows to keep the compatibility among phylogenetic trees belonging to different phylomes. Each organism (represented by its NCBI taxonomy ID) is encoded by using a three letters code that remains stable across all phylomes. Sequences are in the format of the three letters species code plus an unique number (i.e. Hsa0000001), which maintains both the reference to the organism and the reference to the source proteome. Thus, a sequence starting by "Hsa" will always refer to the *Homo sapiens* species, independently of the phylome version and scope, whereas, different sequence numbers can refer to the same protein taken from different proteome versions. For instance, Sce0000185 and Sce0006786 points to the same gene, YBL058W, in two different releases of the yeast proteome. In any case, a correspondence between phylomeDB identifiers and most popular gene and protein names is internally maintained to allow flexible searches and the easy interpretation of phylogenies. Although the web interface of phylomeDB is quite intuitive, a documentation page is available. Technical information about the database structure and web design can be found in the Material and Methods section.

8.1.2 Searching the database

Searching by ID

Available phylogenetic trees and multiple alignments can be accessed by the name of the gene that was used as seed sequence for the phylogenetic pipeline (the one used as a query in the search for putative homologs). By default, ID queries perform searches on all the available phylomes, but searches can be conveniently limited to a specific phylome. Supported IDs vary depending on the source of the seed proteomes, but Uniprot/Swissprot, Ensembl, and the original proteome identifiers are generally valid. Any ID query always results in a list of resources bars for the matching ID. Each resource bar represents the available data for a gene in a given phylome. It is composed of several tree icons (one per every evolutionary model or phylogenetic method tested), two alignments icons (clean an raw alignments) and optional links to any other extra information i.e. orthology

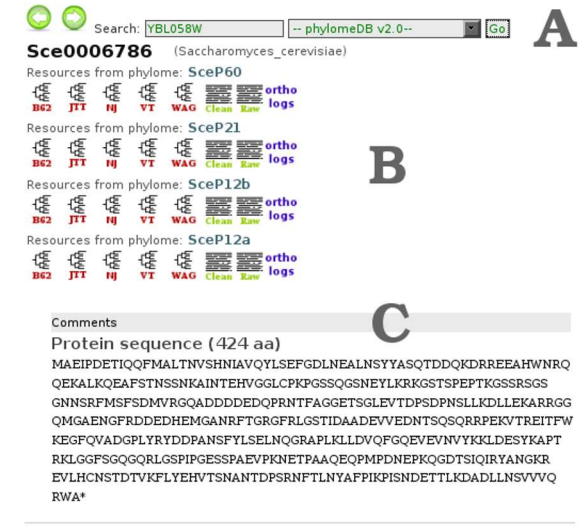


Figure 8.1: Results for an ID search in phylomeDB. A) quick search bar B) Panel showing the available data for the protein. Each row corresponds to the resources available from an independent phylome. In the example, the sequence YBL058W is present in 4 yeast phylomes with a different species coverage. C) Amino acid sequence linked to the ID.

and paralogy prediction. When a gene is present in several phylomes, a resource bar per phylome is shown (Figure 8.1)

Searching by sequence

Alternatively, a similarity based search can be performed to find the most similar sequence hosted in the database. This is internally done by means of a BLAST comparison between the query sequence and the seed proteomes available in phylomeDB. A list of significant hits and links to their available resources are shown.

8.1.3 Linking PhylomeDB from external resources

PhylomeDB resources are easily linkable from external sources such as other web databases. Allowed external queries are shown in table 8.1.

8.1.4 Interactive tree and alignment visualization

Phylogenetic trees

PhylomeDB uses the ETE libraries (see 8.2) to manage phylogenetic trees. Based on such libraries, an interactive tree viewer is available within the web interface that allows phylogenies to be displayed and manipulated on the fly. A set of

query (GET method)	Description
?seqid=ENSP00000347111.	Generates the resources page for the requested sequence id
?seqid=ENSP00000347111&phylomeid=1	Generates the resources page for the provided sequence id in the phylome with the code 1.
?seqid=ENSP00000347111&phylomeid=1&data=tree_JTT	Generates a web page showing the specific requested phylogeny. Valid tree names are in the format "tree_" plus the name of the evolutionary model or phylogenetic method used for its reconstruction. For instance, "tree_JTT", "tree_Blosum62" or "tree_MrBayes" are valid names for phylomes that used such methods in their phylogenetic pipeline.
?seqid=ENSP00000347111&phylomeid=1&data=alg_raw	Generates a web page showing the raw or clean alignment.

Table 8.1: External queries allowed by the phylomeDB web site. Each query type is referred to the following phylomeDB URL: http://phylomedb.bioinfo.cipf.es/find, e.g. *http://phylomedb.bioinfo.cipf.es/find?seqid=ENSP00000347111.*

common features such as rooting, searching or zooming are provided. When a phylogenetic tree is selected, its visualization starts automatically. The image can be downloaded at any time as an standard PNG image. Additionally, when orthology and paralogy predictions are looked up for a gene, a color-coded phylogeny is displayed that describes the evolutionary events (duplications and speciations) in which orthology predictions are based (See screenshot 8.5 at the end of this chapter)

Alignments

Two types of multiple sequences alignment are hosted in PhylomeDB. Raw alignments are those that multiple alignment programs directly produces. Clean alignments, in contrast, are the result of trimming raw alignments to eliminate poorly aligned regions (see methods). Although different trimming criteria are applied in the different phylomes, alignment processing generally includes some kind of gap trimming steps. Alignments of current hosted phylomes has been all processed by using the trimAl program . Inline visualization of alignments is implemented by using the Jalview plug-in (Figure 8.2) Clamp *et al.* (2004) , which is a java application enabling fast viewing of large multiple sequence alignments.

8.2 ETE: a python programming Environment for Tree Exploration

ETE is a framework that provides a set of programming libraries to analyze and manipulate phylogenetic trees (or any other type of hierarchical trees). It implements, among other features, two orthology detection algorithms, a tree reconciliation method, several rooting functions and a programming API to remotely exploit the phylomeDB database. Hierarchical trees are commonly used to represent many types of data, including results from microarray clustering analyses and phylogenetic reconstructions. In recent years their use has been

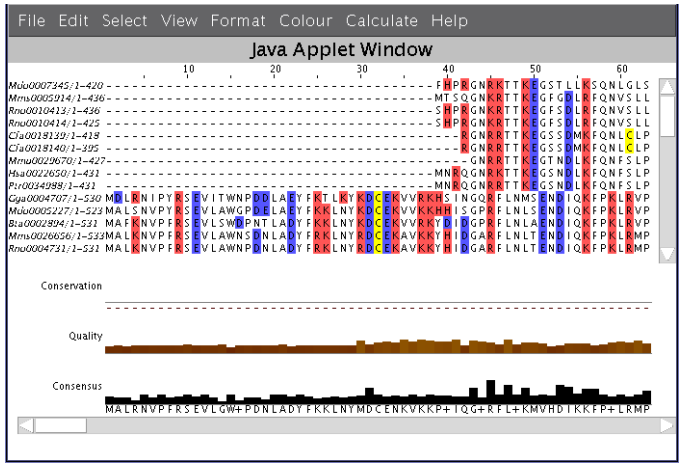


Figure 8.2: A multiple sequence alignment interactively shown on the phylomeDB web interface through the program JalView Clamp *et al.* (2004).

popularized, triggering the need for new software and algorithms. Currently, there are a number of computational tools that assist in the visualization and analysis of trees . Most of these applications are centered around the visualization of the trees and, only in some cases, also include some editing options. Other tools, in contrast, are focused on the implementation of specific analyses but do not provide many possibilities to manipulate the trees. In general, different programs are designed to work with a specific type of data and they only perform a limited set of analyses. As a result, most researchers aiming to perform elaborate analyses have to combine different programs or to implement specific scripts. ETE has been written under a distinct philosophy. To allow for a high degree of flexibility, it is implemented as a collection of python libraries that can be extended, combined and used from third programs. Current version of ETE includes five main extensions: tree management, tree visualization, phylogenetic extension, microarray clustering extension, and the phylomeDB API. Currently, ETE libraries are being used in projects such as GEPAS Tárraga *et al.* (2008), Phylemon Tárraga *et al.* (2007) and PhylomeDB Huerta-Cepas *et al.* (2007). ETE can be downloaded from: http://bioinfo.cipf.es/downloads/ete/

8.2.1 Tree Management

ETE allows to read tree structures from the three most common formats: New Hampshire (NH), New Hampshire eXtended (NHX) and Nexus (through the BioPython parser). When a tree is loaded, its hierarchical structure is internally encoded as a series of tree nodes instances that are connected following a parent-child relationship. Each node provides, itself, many methods to manipulate its connections (add or remove childs, detach from parent, delete, etc...) and to easily access its topology (get parent, terminal, child, sister or descendant nodes). Search and distance methods are also available, and allow the user to

calculate paths between nodes or find specific nodes within large tree structures. The subtree structure under any internal node can be recovered at any time of the analysis. One advantage of ETE is that allows the possibility of adding extra information to the nodes. These data can be used internally or be automatically incorporated, using the NHX format, to the final newick representation.

8.2.2 Tree Visualization

The treeview extension provides a highly programmable drawing system to render a hierarchical tree structure into a custom image. Although a number of predefined visualization modes are distributed with the default installation, custom styles can be easily created from scratch. To do so, ETE makes use of three main concepts (styles, faces and layouts), that allow the user to compose custom tree pictures designs: A 'style' defines the general aspect of a tree node. 'faces' are small drawings (representing, for instance, any node's extra information) that are drawn next to nodes. Finally, 'layouts' are small python functions that control the styles and faces that are applied to each tree node. By combining this features, graphical nodes representation can be dynamically controlled by custom criteria. Several examples and a detailed user manual for the programmable drawing system are available at http://bioinfo.cipf.es/downloads/ete/.

Drawing engine and the visualization interface are based on the new Qt4 open source libraries, which make possible to visualize trees that exceed the max allowed image file dimensions. Performance, however, will depend on each computer resources.

8.2.3 Phylogenetic Extension

Phylogenetic trees are the result of most evolutionary analyses. The phylogenetic extension implements many operations that are common in the analysis of evolutionary trees. Among other features, it allows to link phylogenetic trees to multiple sequence alignments; species names and taxonomic information can be encoded for every leaf node; and midpoint or most-distant-species outgroup can be automatically calculated. Furthermore, two evolutionary event detection methods are provided: One implements the algorithm described in chapter 5, which is based on the species overlap between clades and thus does not depend on the availability of a species tree. The second one, which requires the comparison between the gene tree and a previously defined species tree, implements a strict tree reconciliation algorithm Page and Charleston (1997). Both methods return a list of the detected evolutionary events and the derived orthology/paralogy predictions. This functionality is also fully integrated with the visualization extension, and results can be explored graphically (Figure 8.3)

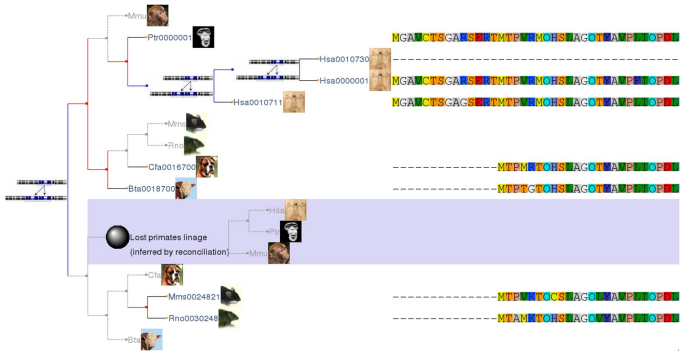


Figure 8.3: A phylogenetic tree visualized with ETE. The phylogeny was obtained by applying the tree reconciliation algorithm implemented in ETE. Dashed lines represent the inferred gene losses, blue lines represent duplication events, and red lines represent speciation events. Sequences associated to each tree leaf are also shown.

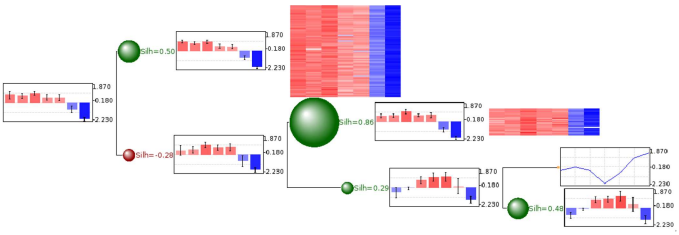


Figure 8.4: Part of a clustering result visualized with ETE. The tree was obtained from a hierarchical clustering analysis on a microarray experiment. Bar plots represent the gene expression profiles grouped by each node. Green and red circles represent the validation analysis of each node performed by their silhouette index calculation.

8.2.4 Microarray Clustering Extension

Microarray expression data is usually encoded as a matrix in which each row represents a gene expression profile across different conditions (columns). A variety of clustering analyses are used to group genes that respond co-ordinately to a given set of conditions, or to group conditions according to their gene expression similarities. In the particular case of hierarchical clustering, clusters are connected in the form of a hierarchical tree. In such trees, genes are represented by terminal nodes whereas internal nodes define the different nested clusters. ETE's clustering extension allows to import microarray matrix files and link them to their associated clustering analysis. Once trees are loaded, expression profiles are accessible from genes (tree leafs) and mean expression patterns are accessible from clusters (internal nodes). Clustering validation techniques can be applied to calculate the inter- and intra-cluster distances, variance or Silhouette width of any partition. Furthermore, predefined node faces an layouts are supplied to visualize clusters together with their expression profiles and to explore the resulting validation analysis (Figure 8.4).

8.2.5 PhylomeDB Extension

ETE 's phylomeDB extension provides a programming API to access the database. It allows to search for specific data and download them for its analysis. When a phylogenetic tree is retrieved, it is automatically linked to the ETE's phylogenetic extension, thus enabling its visualization and analysis within the ETE's framework.

8.2.6 ETE as an standalone application

Finally, ETE can be used as a standalone tree viewer. An executable script implementing most common features of the ETE library can be found within the standard installation. The program can be used to visualize generic, phylogenetic and clustering trees using a set of predefined layouts.

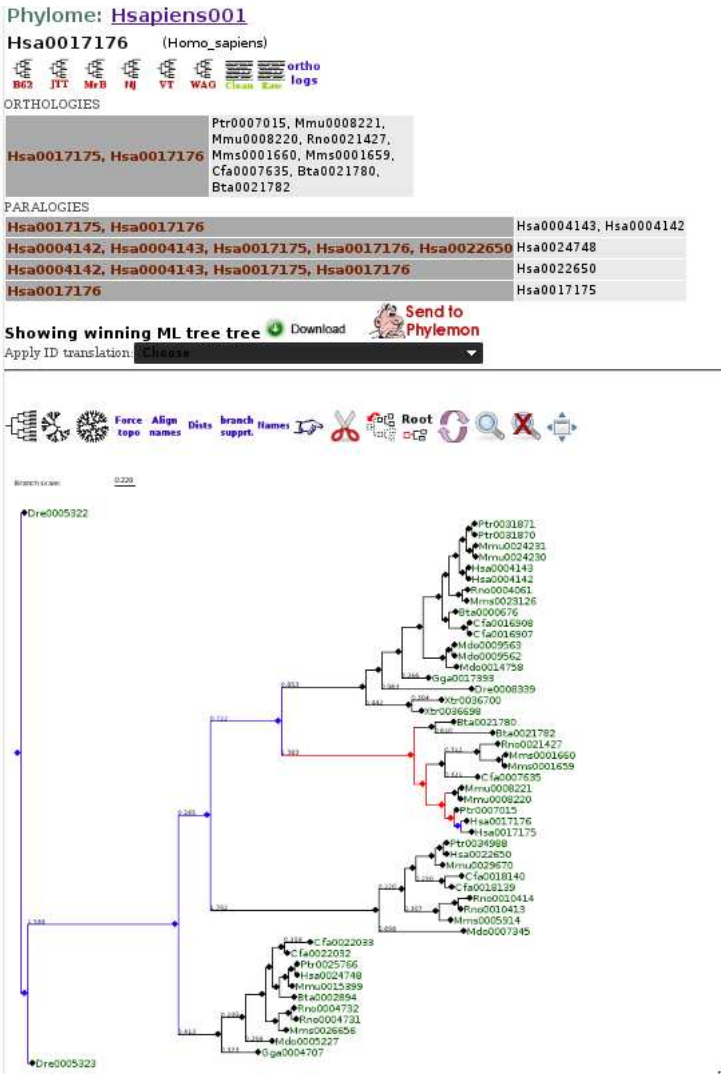


Figure 8.5: A phylomeDB orthology prediction page.

Part III

Materials and Methods

Chapter 9

Biological data sets

9.1 Eukaryotic proteomes

Proteomes derived from 39 fully sequenced eukaryotic genomes (Table 1) were downloaded from Ensembl v36 Birney *et al.* (2006)and the Integr8 database at EBI Pruess *et al.* (2005), except those of *Candida albicans N. crassa* and *C. reinhardtii* , which were obtained from their specific databases. Wherever mitochondrial proteins were not included in the gene set per species, these were downloaded separately from the NCBI eukaryotic organelles site. Mitochondrial genomes from *Caenorhabditis briggsae*, *Giberella zeae*, *Debaromyces hansenii* and *Leishmania major* have apparently not been deposited in the public databases and, therefore, are missing from this study. The final proteome database contains 542,423 unique protein sequences from 39 different genomes (Table 5.1).

9.2 Synteny based yeast duplicates

Sacaromices cerevisiae paralogous genes originated from the same whole genome duplication were obtained from the synteny based predictions from the Yeast Gene Order Browser web server (http://wolfe.gen.tcd.ie/ygob/). This dataset is described in Byrne and Wolfe (2005) and comprises a total of 450 pairs of paralogous genes. For the analysis in chapter 6, we derived a filtered list containing only those 250 paralog pairs that were present in the 3 post-WGD species considered: *S.castellii* and *C.glabrata* and *S.cerevisiae*.

9.3 Human and Mouse expression data sets

Expression data of 79 human and 59 mouse healthy tissues were retrieved from the SymAtlas project at the Genomics Institute of the Novartis Research Foundation (GNF) (http://symatlas.gnf.org/SymAtlas/). We used the two-samples

Affymetrix HG-U133A probeset with MAS5 normalization for human and the GNF1M data set with MAS5 normalization for mouse. To determine whether a gene was expressed or not in a given tissue, we used the data sets including Presence and Absence calls. A gene was considered to be present in a tissue if Presence calls were reported for the two samples included in the probeset. We assumed marginal expression (M) to be Absences. Probes mapping to several genes were removed from the analysis.

9.4 PhylomeDB database

PhylomeDB has been built using a relational database implemented with MySQL5.0 (http://www.mysql.com/). Official web site (htttp://phylomedb.bioinfo.cipf.es) uses basic HTML and Asynchronous JavaScript And XML (AJAX) features. Pre-release area of phylomeDB (htttp://beta-phylomedb.bioinfo.cipf.es) is implemented using a modified version of the dokuwiki (http://www.dokuwiki.org/) software. Phylogenetic tree management and visualization is performed by the Environment of Tree Exploration software (Chapter 8). Visualization of multiple sequence alingments is implemented by the program Jalview Clamp *et al.* (2004). Although most of modern browsers are supported, Firefox is recommend for a full compatibility.

Chapter 10

Comparative genomics and phylogenetic methods

10.1 Sequence similarity searches

For each protein sequence, a Smith-Waterman Smith and Waterman (1981) search was performed against a blast database containing the proteomes of all 39 species considered for the reconstruction of the human phylome. In order to exclude isoforms from the subsequent analysis, blast database did only contain the larger protein isoform from each gene. Only matches with a significant similarity (E-value $<$ 10-3) that aligned with a continuous region longer than 50% of the query sequence were selected.

10.2 Multiple sequence alignment and phylogenetic reconstructions

The sets of homologous protein sequences were aligned using MUSCLE 3.6 Edgar (2004). Positions in the alignment with gaps in more than 10% of the sequences were eliminated before the phylogenetic analysis, unless this procedure removed more than one-third of the positions in the alignment. In such cases the percentage of sequences with gaps allowed was automatically increased until at least two-thirds of the initial positions were conserved.

NJ trees were derived using scoredist distances as implemented in BioNJ Gascuel (1997). ML trees were inferred from the alignments using PhyML v2.4.4 Guindon and Gascuel (2003). For each protein family ML trees were reconstructed using four different evolutionary models (JTT, WAG, BLOSUM62 and VT), except for the 13 mitochondrial encoded proteins in which the mtREV model was also used. In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was used. The gamma parameter and the

fraction of invariant positions were estimated from the data. The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the AIC criterion . In order to obtain support values of all tree partitions, the ML tree produced by the best-fitting model was used as a seed for a Bayesian analysis by running MrBayes Ronquist and Huelsenbeck (2003) for 100,000 generations in two runs of two chains each,and allowing branch swapping and re-estimation of the gamma distribution parameters. The posterior probability of each tree partition was estimated by sampling the trees every 100 generations after discarding the first 25%. This approach to obtain support values was faster than performing standard bootstrap analysis with PhyML. Starting the MrBayes runs with an already optimized tree resulted in fair levels of convergence being reached after fewer generations than in standard MrBayes analyses. A final tree produced by this Bayesian reconstruction consists of a consensus phylogeny, using the 'halfcompat' option in MrBayes, in which partitions with a posterior probability lower than 0.5 are collapsed. Unless stated otherwise the consensus tree produced by MrBayes analysis was used in all analyses.

10.3 dS ratio between paralogs genes

Pairwise sequence comparisons were performed by using the program codeml within the PAML package Yang (1997). Each pair of paralogous proteins were aligned using the program muscle with default parameters. Pairwise alignments were performed using amino acid sequeces and the program MUSCLE v3.6 Edgar (2004). Subsequently, each position of the amino acid alignment was mapped to its corresponding codon in the nucleotide sequence. Alignment positions containing gaps were discarded.

Chapter 11

Analytic methods and algorithms

11.1 Detection of evolutionary events

We used a phylogeny-based algorithm to detect duplication and speciation events on the trees. In contrast to alternative phylogeny-based methods that use reconciliation of the gene tree with the species tree to infer duplication events, our approach does not require any previous fully resolved species topology. The only evolutionary information required is that used to root the trees to define a polarity so each internal node is connected to two children nodes.

The orthology prediction algorithm was run independently for each human gene using the its primary tree (the one reconstructed using such sequence as a seed gene). The algorithm was implemented in a series of python scripts specifically developed for this project and now part of the ETE library (8). To map duplication and speciation events on an internal node of the tree, the algorithm proceeds as follows. First, two tree partitions are defined that contain the sequences connected to each of the two children nodes. Second, a species-overlap score is defined between the two partitions as follows: species common to both partitions/species in any of the partitions. Third, if the score is higher than a given threshold the node is mapped as a duplication event, otherwise it is considered a speciation event. In the present study the species-overlap threshold was set to 0.0 - that is, no common species between the two partitions were allowed - because this produced the best results in the benchmark. The algorithm does so for all internal nodes in the tree. Once all the nodes in the tree are marked as a duplication or speciation event, the algorithm establishes orthology relationships between the seed protein and other proteins in the tree. For each protein, the algorithm tracks the nodes that connect it to the seed protein and establishes an orthology relationship only if this connection proceeds exclusively through speciation nodes, disregarding intra-specific duplications. After mapping speciation and duplication nodes onto the phylogeny, several situations may arise in which

orthology relationships are not one-to-one relationships, but rather one-to-many or many-to-many. To root the trees the following procedure was used. The species present in the tree were grouped according to the branching pattern of the tree in Figure 3; thus, non-opisthokont species constitute the deepest group, followed by 'fungi', 'other metazoans', 'urochordates', and so on. Among the sequences belonging to the deepest group with representatives in the tree, the one with the longest distance to the seed protein was chosen as the out-group.

11.2 Topology scanning algorithm

The algorithm used here to search for specific topologies within the phylome is described elsewhere Gabaldón and Huynen (2003). In brief, from an un-rooted tree the algorithm generates all possible partitions that contain the seed sequence. That is, the algorithm proceeds sequentially throughout all internal edges of the tree. At each internal edge it generates two partitions, of which only one contains the seed sequence. The species represented in each such partition are tracked and those trees with a partition fulfilling a set of rules defined by the user are selected. The set of rules defined by the user are defined as a set of species that are allowed in a partition, and rules can be combined so that specific evolutionary scenarios are defined. For instance, a partition supporting the grouping of rodents and primates to the exclusion of laurasatherians can be defined as a partition containing any sequence (s) from primates (*Homo sapiens*, *Macacca mulata*, *Pan troglodites*) and any sequence (s) from rodents (*Mus musculus*, *Rattus norvergicus*) within a larger partition that contains these sequence plus any sequence (s) from Laurasatherians (*Canis familiaris*, *Bos taurus*). Sequences from other species are not allowed in the partition and the presence of the seed sequence in the partition is required. This algorithm has been implemented in a series of Python scripts developed for this project. In the topology scanning analyses presented here we discarded the trees based on alignments in which less than 100 columns were left after applying the gap filter. This procedure eliminated 1,714 (7.9%) from the total phylome.

11.3 Orthology prediction benchmarking

The reference set used in a recent benchmark of orthology assignment methods Hulsen *et al.* (2006) was used to compute the number of true positives (TPs), false positives (FPs) and false negatives (FNs) yielded by each method. For each method the sensitivity, S = TP/(TP + FN), and the positive predictive value, P = TP/(TP + FP), were computed.

11.4 Algorithm for the topological dating

Duplication events detected by the topology scanning algorithm can be assigned to different evolutionary periods by examining the species represented after the duplication event. To do so we used as a reference a set of clearly defined phylogenetic relationships that mark the major branching points in the lineage leading to hominids (Figure 3). For each duplication event, all the species represented after the duplication node are tracked and the duplication is assigned to the deepest branching point in the reference tree that contains all these species. For instance, if only sequences from mammals and fishes are found after the duplication event, this duplication is assigned to the branching point that is at the base of vertebrates. The scanning algorithm can detect only duplications that occurred after the root of the tree; for example, if a tree is rooted in a fungal sequence, only duplications that occurred in the metazoan lineage could be detected. Therefore, to compare the results obtained at the different evolutionary stages we computed the relative number of duplication events per gene at each branching point. This was done by dividing the number of duplication events mapped at a particular evolutionary stage by the number of trees rooted at a deeper branching point; for example, duplications that occurred at the base of metazoans were divided by the number of trees rooted on either a fungal or a non-opisthokont sequence.

11.5 Tissue expression complementarity score

The degree of tissue expression complementarity between the expression profiles of two sets of paralogous genes was calculated by measuring the relative number of tissues in which only one set but not the other was expressed over the total number of tissues with expression data. Let p_a be the gene expression pattern of gene *'a'*; d_{a-b} the number of tissues in which gene *'a'* is differentially expressed in respect to a second gene *'b'*; and t_a the total number of tissues in which gene a is expressed. Hence, the degree of tissue expression complementarity (C) between genes *'a'* and *'b'* is:

$$C = \frac{\frac{d1}{t1} + \frac{d2}{t2}}{2}$$

11.6 Expression breadth

We used two different measures to quantify the amplitude of the expression of a paralogous gene set: the goblal and mean expression breadth. We consider global expression breadth as the total number of tissues in which at least one gene of a paralogous set is expressed. By contrast, the mean expression breadth is the average number of tissues in which all genes within the same paralogous group are expressed.

Part IV

Discussion and Concluding remarks

Chapter 12

Summarizing discussion

This thesis comprises a series of studies on the evolution of the human genome, which are performed through the exploitation of large collections of gene phylogenies (phylomes). In this section, we will discuss the main results obtained in this thesis. Some of these are purely methodological, but we consider them to be useful for future studies; other have a deeper theoretical relevance and we intend to contribute to current discussions attending to our own experience.

12.1 Meeting the challenge of reconstructing high-quality phylomes.

Previous studies have addressed the reconstruction of complete phylomes (Sicheritz-Pontén and Andersson, 2001; Gabaldón and Huynen, 2005). However, the species coverage, and the use of sophisticated phylogenetic pipelines have been traditionaly limited by computational constrains. For this reason, most previous attempts have focused on bacterial genomes (comprising a significantly lower number of genes than eukaryotes) and less accurate methods to infer phylogenies, such as NJ.

Fortunately, current improvements in algorithms and computer capabilities have paved the way for the application of more sophisticated methods at genomic scale. In this respect, several remarkable efforts exist that provide accurate phylogenomic resources. For example, the Ensembl project provides, from version 41, a Maximum Likelihood reconstructed phylogeny for each of their predicted gene families. Similarly, TreeFam (an Ensembl related project), extends the same methodology for a larger number of animal species. Interestingly, TreeFam provides also a subset of manually reviewed phylogenies.

It can be noticed, however, that many restrictions are still present in the current phylogenomic approaches. For instance, neither of mentioned projects include evolutionary model testing or alignment pre-processing steps in their

pipelines. By contrast, the same evolutionary hypothesis are equally assumed for the evolution of all genes, something that has been shown to lead into errors . Additionally, in order to mitigate the high computational cost, a family-based approach is usually adopted to reduce the number of trees to be reconstructed (see 12.2).

In chapter 5 of the present thesis we have shown the feasibility of applying a sophisticated pipeline to the reconstruction of large collection of phylogenies by compiling the first version of the human phylome, which comprises more than 100,000 phylogenetic trees and 20,000 MSAs. Besides the usefulness derived from such data (farther discuss in next sections bellow), our methodology has illustrated that the most accurate phylogenetic methods can currently be applied, although at a high computational cost, to genome-wide studies. To our knowledge, the time invested in generating the human phylome represents the largest computational effort to reconstruct the evolutionary history of all human genes.

The importance of high quality approaches in molecular phylogeny has been addressed many times . One of the key issues in this respect is that the evolutionary model should not be equally assumed for all gene phylogenies. Our pipeline tested up to 5 protein evolutionary matrices on more than 20,000 alignments and found an heterogeneous distribution of models among different gene families, proving that the general assumption of a fixed model blocks the finding of the best phylogenetic tree. Interestingly, the most representative model found in our analysis, JTT (see 5.2), does not match with the one assumed in others projects.

Another important question concerning any large-scale analysis is the cost-effectiveness of the approach. There is no doubt that the manual curation of the results by human experts is the best approach to interpret and refine the results; however, it is also evident that such a methodology cannot account for the exponential grow of current sequence and genome databases. In this respect, the set of methods developed in this thesis form a fully automatic pipeline that includes also the automatic extraction of relevant information from large collection of trees. We believe this to represent an important advantage in the genome era.

However, it must be noted that our approach needs a level of computer power that is still far from the reach of standard desktop computers. Running our pipeline over all protein-coding genes in the human genome requires massive computational resources including the Mare Nostrum supercomputer of the Barcelona Supercomputing center (up 2 months of continuous parallel computing on ~200 modern processors).. Our current projects, which are populating the PhylomeDB database are running on a cluster of 200 64 bits CPU at CIPF and in the Mare Nostrum supercomputer through several approved projects.

12.2 Gene-based versus family-based approaches

Two main approaches can be considered to reconstruct the evolutionary history of all genes of an organism, these are, namely gene-based or family-based. In the

former , idependent phylogenies are inferred for every single gene, whereas the latter strategy relies on a previous grouping of sequences into families and the subsequent reconstruction of a single phylogeny per family. The main difference between both methods is that the family-based approach does not produce (or should not produce) redundant information, since each sequence is only present in a single phylogeny. By contrast, gene trees inferred from closely related sequences would generally lead into quite similar or even equal topologies.

At first glance, redundancy may present obvious drawbacks. For instance, the computational cost of the overall analysis will be significantly greater in a gene-based approach than in family-based one , simply because the same data is processed several times. Moreover, the automatic interpretation of gene trees becomes substantially more complex, since it will have to tackle with duplicated or, even worst, unmatching results.

However, despite these drawbacks, redundancy can be certainly useful to to avoid some of the problems associated to the family-based approach. Let us first take a look to the drawbacks associated to a family-based approach. Families of sequences are usually automatically inferred by methods based on all-against-all similarity sequence comparisons . Such comparisons, which form a distance matrix among all sequences, are used to create a network (a graph) in which nodes are the sequences and edges the degree of similarity among them. Subsequently, the network is automatically processed to detect groups of sequences(modules) that are more connected among them than to any other group. These modules are considered gene families. Most of the drawbacks of this approach are associated to the initial presumption that a single phylogenetic tree (and the corresponding MSA) can model the evolution of sometimes very large families, including singular cases of sequences from the same family that are barely similar. The consequence of this are poor alignments and a higher rate of tree artifacts. In addition, the detection of families constitutes an added source of complexity.

Alternatively, the gene-based approach allow us to split big families into many overlapping gene phylogenies, ensuring that all the sequences included in the analysis are significantly similar to the seed gene (the one used to search for homologs). We have found this approach to produce quite reliable results by considering that the most accurate representation of a gene evolutionary history is within its own phylogeny, obviating that the same gene is also present in other trees.

Overall, we think that, although the family-based strategy is accurate enough for most of the necessities, the redundancy associated to gene trees can be exploited to obtain more precise estimates. For example, it would be interesting to evaluate whether redundancy can be applied to provide a level of support to the prediction of orthologs.

12.3 The Tree of Life

The Tree of Life (TOL) is a metaphor coined by Charles Darwin to express the idea that the evolution of all species can be expressed as dicotomic tree with a single common root (ancestor). In such a tree structure, each internal node is considered as an ancestral organism, whereas leafs represent the modern species. Although deciphering the TOL is a major task in biology, many important branch points remain currently unclear . Molecular phylogenetics have contributed to elucidate many species relationships, however, how can phylogenomics contributes to the study of the species relationships?

It exists several approaches that aim to compile the data coming from different phylogenetic sources into a single species tree: namely super-trees, concatenation of sequence alignments, gene content analysis or whole genome trees . Among them, concatenation is probably the most extended one, since in principle, it allows for more robust and representative phylogenies by increasing the number of informative sites in the analysis. The usual procedure consists of performing a series of independent multiple alignments including different sets of genes, concatenate the resulting aligned sequences, and build a larger alignment that is subsequently used to infer a single phylogenetic tree. The inclusion of data from various sources allow us to amplify the global phylogenetic signal and to increase the resolving power, for example, when the signal is masked by homoplasy.

However, in order to concatenate alignments, they ideally have to content an equal number of sequences, that of the number of species. Hence, alignments constructed from larger or smaller sets of homologous sequences are discarded from concatenation. In other words, only those gene families conserved in single copy across all species are susceptible to be used by concatenation methods.Unavoidably, the number of genes meeting such condition decreases as the number of species included grows. To attenuate this effect, genes absent in a few genomes can be included by introducing gaps in the species missing them and methods to select one gene from few recent paralogs can be applied. For instance, in a recently reported tree of life including 191 species, only 31 genes were used . This has raised some criticism as to whether such a reduced number of genes can fairly represent the evolution of whole genomes. Reduced gene sets are subject to biases caused by gene-sampling effects or by differences in lengths, since house-keeping, widespread genes, and specially those with long sequences, contribute with a higher amount of information.

In this thesis we have explored the possibilities that the analysis of complete phylomes (not biased by subsets of genes) may offers to the resolution of the TOL. Our results have revealed that, in spite of a genome-wide perspective, a high variability (even more than expected) is found among the phylogenies of different genes. Indeed, the high level of incongruence found for the three selected scenarios considered in chapter 5(see 5.3) suggests that current phylogenomics will probably fail into solving other uncertainties in the TOL.

However, a recent study has revealed a high level of compatibility between concatenation approaches and the support extrated from the analysis of phylomes

(Marcet-Houben and Gabaldón, 2007). In such work, several fungal phylomes were reconstructed following the general pipeline described in this thesis. Parallel, a concatenated alignment was created from 69 widespread genes presenting a single ortholog across at least58 out of the 60 fungal organisms considered. Authors inferred a species tree from the concatenated analysis using ML, and provided two types of support values for each branch: a parametric bootstrap, and the frequency with which the same monophyletic partition was found on the complete phylome landscape (phylome support). Interestingly, although phylome support yielded the same high levels of topological visibilities, the most represented topology in the phylome landscape was always coincident with the topology inferred by the concatenated alignment. Only in two cases, in which species relationships are largely debated, partitions on the concatenation-tree were not supported by the most represented scenario in the phylome landscape.

12.4 Improving orthology predictions

The importance of orthologs has been discussed in different parts of this thesis 1.2.1, 5.4). An immediate application derived from the availability of a complete phylome is the possibility of deriving orthology and paralogy relationships as inferred from the gene phylogenies. We addressed the genome-wide identification of orthology and provided a full catalog of the human orthology and paralogy relationships on 38 species (5.4). Our comparisons have also shown the phylome-based predictions to overcome the accuracy of the most extended methodologies when they are applied on intricate cases (Figure 5.3). Interestingly, among the three phylogenetic methods included in the comparison, only the phylogenetic trees reconstructed following statistical approaches (Ensembl used ML and Phylome predictions used ML and Bayesian) were able to overcome the similarity based methods. By contrast, phylogenetic trees reconstructed using NJ (PGT in Figure), produced an high number of false positives, thus suggesting that the quality of the phylogeny is crucial to obtain reliable results.

We have previously discussed the importance that prediction methods could become fully automatic. In this respect, our implementation of the species-overlap algorithm have shown very good levels of accuracy and tolerance to phylogenetic artifacts as compared to other alternatives. For example, in a recent study [Marcet-Houben and Gabaldón] , the accuracy of the species-overlap strategy and that from the strict reconciliation methods have been evaluated for a set of fungal orthology relationships, finding that sensitivity of the former is much lower than the species-overlap (See Figure 12.1).

In summary, we can conclude that the combination of high quality phylomes and automatic interpretation of phylogenies provides a valuable framework to increase the resolution of current orthology predictions (for example by distinguishing among many-to-many and one-to-many relationships) and for its application to non-model or new sequenced organisms.

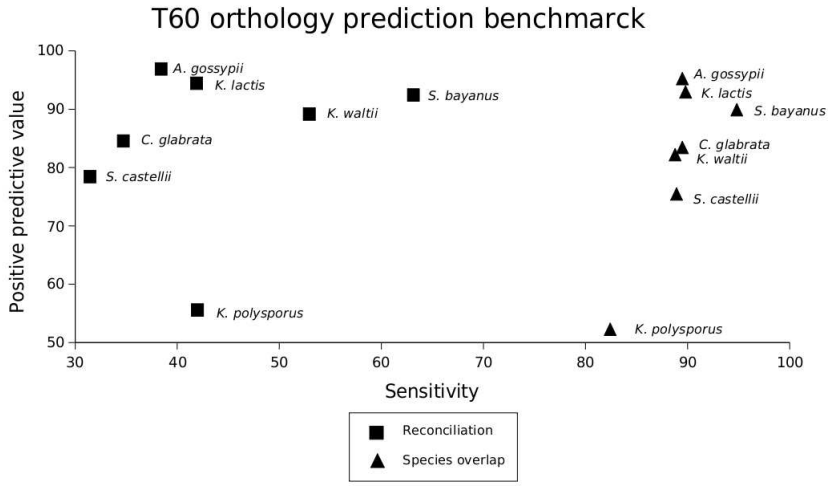


Figure 12.1: Comparison of two phylogeny-based methods for orthology detection: "strict reconciliation" and "species overlap". Species overlap predictions were performed by following the algorithm presented in this thesis and implemented in ETE (Chapter 8). Analysis and Figure obtained from (Marcet-Houben and Gabaldón, 2007).

12.5 Studying the evolution of gene expression

Several methological differences can be found between the analyses presented in chapter 5 and previous studies on the same area.

First, the use phylogenetic methods provides a broader perspective on the gene duplication process, allowing, for example, to identify many-to-many paralogy relationships rather only one-to-one. The fact of considering multiple paralogs under the same duplication events allow us to consider more complex scenarios of gene expression variants.

Another aspect, discussed also in chapter 6, is the method used to estimate the age of the duplications events. We have found that dS ratios (mostly used in similar studies) can present many drawbacks when they are used as proxies to the divergence time. Again, we think that a phylogenetic view can better estimate the relative time in which a duplication or speciation took place in the evolution. Our method have considered 9 main periods in the evolution of eukaryotes to perform the analysis.

The degree of variability associated with microarray experiments could cause the over or under estimation of expression differences found between genes. In order to minimize such effects, we used two data sets coming from the same laboratory and that were performed under the same conditions and platforms. In addition, all our analysis have only focused on the presence or absence expression profiles of genes across tissues, discarding the differences among the levels of expression. We consider this a more conservative way that accounts only for the more evident differences.

Another important aspect taken into account is the possible biased introduced when comparing two different data sets . In this respect, the analysis requiring the comparisons between inter-species and intra-species estimation of expression differences, were redefined to include only a single type comparisons. For example, in the case of the orthologs versus paralogs comparison, the tissue expression complementarity of paralogs (intra-species estimates) was calculated using the intra-species paralogy relationships. In other words, considering as paralog the mouse ortholog of the actual human paralog. Furthermore, differences between orthologs and paralogs were calculated in terms of their distribution of expression complementarity values rather than in absolute terms.

12.6 A model for the evolution of tissue expression and its implications for gene retention after duplication

Expression similarity among duplicated genes, measured as the correlation coefficient between their transcription patterns, has been reported to decrease linearly as the divergence time increases . Such a linear correlation implies that expression differences have been accumulated during the course of evolution in a clockwise fashion , somehow suggesting a neutral genetic drift on the gene expression pattern of genes . This idea has attracted considerable attention and several neutralistic models have been indeed developed. However, if we consider that variations on gene expression patterns can be regarded as functional changes, the expression divergence among paralogs can be better understood under the sub and neo-functionalization hypotheses. According to such models, the retention of duplicated genes would come from the acquisition of novel transcription patterns or, in the case of sub-functionalization, from complementary patterns that render both duplicates indispensable to perform the original function. In either case, the differential expression profile would be subjected to positive selection and the retention of both genes would be favored.

Considering this, how does this model fit with the linear increase of expression changes observed among paralogs? We postulate that not all expression divergence is actually accounting for the retention of duplicates. Thus, while most of transcriptional changes might be under neutral evolution, only part of them would be under positive selection. We have isolated the especial case of expression divergence that leads into complementary spatial patterns among duplicates, which represents a clear scenario favoring their retention, and we have found no linear correlation between the age of duplicates and the level of tissue expression complementarity. By contrast, similar levels are found among most of the periods considered. Only duplicates associated to very old or very recent periods showed distinct distributions of expression complementarity. Interestingly, when the same analysis is repeated over the gene expression similarity rather than complementarity, we reproduce the previous observations of linear correlation.

In summary, we propose that a mixed model could be operating on the evolu-

tion of the transcriptome. First, the surprising level of tissue expression complementarity we have found between orthologs suggests that changes in the expression pattern of genes are indeed expected to occur in a rather frequent fashion during the course of evolution. Such changes can be the result of mutations in the promoter regions or in any regulatory factor affecting its pattern of the gene expression, and they would occur in a (mostly) random way.

After the duplication of a gene, the resulting duplicates will evolve independently in terms of their pattern of tissue expression (just like any other gene), but they will be initially have more chance of loosing one the redundant copies. The inherent expression divergence assumed by the model can eventually lead to the situation in which the original and necessary expression pattern is only possible by the combination of both duplicates (sub-specialization) or that a new and favorable expression pattern is acquired by one of the duplicates (neo-functionalization). Note that this holds even if both duplicates perform exactly the same function and even if their coding sequences are identical, since the mere differences on gene expression patterns may constitute a factor to conserve both duplicates. In such a case, the retention of both duplicates and their complementary pattern will be favored by the evolution. This would explain the observed higher degree of complementarity observed between paralogs as compared to orthologs. Furthermore, another result that supports this idea is the fact that many of the complementarity found between human-mouse paralogs were likely acquired prior to the divergence of these two lineages.

That the process of tissue expression divergence has likely facilitated the retention of duplicated genes does not necessarily imply that all retained duplicates might have undergone such process, or that the complementarity observed nowadays is the one that favored its retention. The random drift initially assumed can indeed account for much variability across the evolution of genes. We found, for example, many cases in which the observed specificity cannot explain the retention of the duplicated genes, simply because the duplication is much more ancient than the origin of the specific tissue itself. Our model predicts that many of the tissue-expression complementarity that facilitated in the past the retention of duplicates have likely been shaped during the course of evolution.

12.7 Future perspectives on the use of phylomes

An attractive feature of the methods presented in this thesis is their high level of customization and extensibility. This can be nicely illustrated by the fact that, only a year after the publication of the human phylome (Huerta-Cepas *et al.*, 2007), similar pipelines have been successfully applied to other organisms and scenarios. Some of such phylomes respond to our own research interests and others have been instigated through collaborations with other groups. For example, the methods presented here are being successfully applied to the detection of Horizontal Gene Transfer, orthology prediction on non-model species or functional annotation of new organisms. In addition, many of these projects use PhylomeDB to allocate their results. Table 12.1 shows a list of new phylomes that are also

being integrated in phylomeDB for its public use.

species name	sp. range	models	status
Homo sapiens	39	JTT, WAG, B62, VT, MtRev	Huerta-Cepas, et al 2007
Sacaromyces cerevisiae	60	JTT, WAG, B62, VT	submited
Sacaromyces cerevisiae	21	JTT, WAG, B62, VT	submited
Sacaromyces cerevisiae	12	JTT, WAG, B62, VT	submited
Sacaromyces cerevisiae	12	JTT, WAG, B62, VT	submited
Candida glabrata	60	JTT, WAG, B62, VT	in preparation
Drosophila melanogaster	24	JTT, WAG, B62, VT	in preparation
Candida albicans	60	JTT,WAG, B62, VT	in preparation
Aspergillus terreus	60	JTT,WAG, B62, VT	in preparation
Aspergillus gossipy	60	JTT,WAG, B62, VT	in preparation
Magnaporthe grisea	?	JTT,WAG, B62, VT	in preparation
Salinibacter ruber	426	JTT, WAG, VT	submited
Salinibacter M8	426	JTT, WAG, VT	submited

Table 12.1: List of phylomes currently (September 10, 2008) hosted in phylomeDB. 'species name' is the organism whose gene was used as seed in the phylogenetic pipeline; 'sp. range' is the number of species considered in the analysis; 'trees' is the total number of phylogenies reconstructed; 'models' are the evolutionary models tested by the pipeline; and 'status' is the current status (published, in preparation, submitted) of the phylome analysis.

Comparative phylogenomics

The possibility of reconstructing whole phylomes from complete genomes facilitates the customary use of phylogenetic methods in the framework of comparative genomics. This is extremely useful when a comparative study aims to help in the characterization of a recently sequenced genome. Through a collaboration with Dr. Josefa Anton (University of Alicante), we have been involved in the characterization of a new strain (named M8) of the *Salinibacter ruber* species, an organism adapted to hypersaline environments. This organism shares its habitat with extremely halophilic Archaea species, providing a suitable scenario for an inter domain Horizontal Gene Transfer (HGT). The genomic analysis of the unique known strain of S.ruber, M31, suggested that this was indeed the case, although at a more modest frequency as initially expected . The sequencing of a new genome from a distinct subspecies allow us to investigate differences on such process at a more detailed level. To do so, we reconstructed both subspecies phylomes (M8 and M31) in a context of 426 bacterial and archaeal organisms. Subsequently, the same methodology described was applied on phylomes, detecting a total of 40 candidates well supported by the phylogenetic analysis. Interestingly, 6 of such cases were specific to the new strain (M8), while 18 from the previously reported in M31 were absent in M8. Some of the candidates were indeed clearly identified as archaeal specific genes. Is the case of the gene coding for halorhodopsin, a retinal chloride pump that had not been described so far in any bacterial or eukaryal genome.

Phylomes and the species overlap algorithm served also to identify 25 duplication occurred in the M8 but not in M31, as well as M31 intra-specific. The

same analysis was used to establish all the orthology relationships among both strains and to identify their strain-specific genes.

Phylomes for genome functional annotation pipelines.

A phylome can be useful from the very early stages after a genome sequence have been assembled. We have realized that the use of phylogeny-based orthology predictions in the early functional annotation of sequenced genomes can readily improve the level of annotation of published genomes. The low rate of false positive predictions achieved by phylogeny-based methods (see benchmarks 5.3, 12.1) makes them especially suited for cases in which orthology prediction is used for the transfer of functional annotations among model species. In such cases, minimizing the level of wrong assignments, which will lead to wrong annotations, is more important than reaching a high coverage at a cost of many false assignments. The phylome-based strategy presented in this thesis is being evaluated for the automatic annotation of some ongoing genome sequencing projects including the pea aphid *A. pisum* and the trematode *Schistosoma mansonii.*

Other applications.

- Continuing in the direction of the analysis of the gene duplication process, we are using a specially suited version of the human phylome to explore the selective pressures occurring across the human linage on duplicated branches. To this end, phylogenetic trees are used to localize the duplication events and paralogs are searched for positive or relaxed selection.
- MANTIS (Tzika *et al.*, 2008) is a database for the evolutionary information for (i) gene gains and losses, (ii) gene content of ancestral organisms and (iii) functional profiles and tissue specific variations. MANTIS uses phylogenetic trees and a maximum-likelihood function to infer such data. The use of the phylogenetic trees reconstructed in the human phylome have provided a substantial improvement in the quality of the results (personal communication).
- PeroxisomeDB (Schlüter *et al.*, 2006) is a database that includes the complete peroxisomal proteome of *Homo sapiens* and *Saccharomyces cerevisiae*, by gathering, updating and integrating the available genetic and functional information on peroxisomal genes. Phylogenetic data was obtained the pipeline presented in this thesis.
- We have reconstructed the *Drosophila Melanogaster* phylome covering all the current fully sequenced *Drosophila* species. This phylome is being used to explore the specific relationships among the genomes of close related species.

Chapter 13

Conclusions

1. We have shown the feasibility of reconstructing complex phylomes, comprising the evolutionary histories of all genes from a given species and their homologs in dozens of other genomes. The pipeline and genome sampling is fully automated and can easily be tailored for specific needs, therefore paving the way for the reconstructions of other phylomes using different parameters or species sampling. Because of its significance, we have initially applied this pipeline to the human genome.

2. The analysis of the Human Phylome reinforce the older view that topological differences among phylogenies of proteins are to be expected even in the absence of HGT and underscore the danger of gene-sampling effects when combining the phylogenetic signals of several genes . We share the view of others that there is an urgent need for improved models of molecular evolution that account for the inherent phylogenetic noise in the protein record and of new genomic characters that are less prone to homoplastic effects.

3. We have developed a flexible algorithm that allows to automatically extract the evolutionary events (speciation and duplication) from large collection of gene phylogenies. Our phylogeny-based approach have shown to overcome current similarity-based methods when they are applied to the identification of intricate orthology relationships. The full catalog of predicted human orthology and paralogy relationships has been also supplied.

4. We have shown the main drawbacks of dS (synonymous substitution rates) used as a time estimator. We have suggested an alternative phylogeny based approach.

5. We have concluded that the evolution gene expression patterns among paralogs might be regarded under a mixed model that accounts for both neutral and non-neutral variations.

6. We have developed a public database containing the phylogenetic trees, multiple sequence alignments, and orthology/paralogy predictions from a number of phylomes (http://phylomedb.bioinfo.cipf.es)

7. We have released ETE, a package of bioinformatic tools and programming libraries for tree management and visualization.

Appendix A

Resumen en castellano

A.1 Secuencias moleculares, filogenia y Bioinformática

La biología molecular es el campo encargado de estudiar la organización y el funcionamiento de los elementos más pequeños que componen la vida, las llamadas (macro) moléculas. Dentro de esta amplia definición se engloban multitud de procesos y elementos que definen, en último termino, la intrincada maquinaria celular. Sin embargo, entre todas estas relaciones, la intima conexión existente entre los genes (ADN) y la síntesis de proteínas destaca especialmente. En esencia, el ADN es la macromolécula que codifica toda la información necesaria para construir las proteínas; y son estas últimas las que componen el catálogo de funciones disponible en una célula.

Ambos tipos de moléculas basan su composición en la polimerización de un número limitado de componentes que, a modo de eslabones, conforman largas cadenas. En el caso de los genes, los eslabones son llamados nucleótidos; mientras que las proteínas están compuestas de amino ácidos. Así, la ordenación específica de nucleótidos o aminoácidos es el único aspecto que distingue un gen (o proteína) de otros. Por este motivo, el conocimiento de su secuencia representa una valiosa fuente de información acerca de su función específica.

La tarea de descubrir la secuencia de un gen (y por tanto de la proteína que codifica) se conoce como "proceso de secuenciación", y las técnicas para llevarlo a cabo han avanzado a pasos agigantados en las últimas décadas. En concreto, el desarrollo de métodos de secuenciación a gran escala ha permitido, desde no más de un par de décadas, determinar ya no la secuencia de unos cuantos genes, sino la del genoma completo de un organismo. Sin duda, estos avances han tenido una gran repercusión en muchas áreas de la biología, ya que los datos generados constituyen la información primaria que distingue a especies e incluso individuos.

Uno de los campos más directamente afectados por esta reciente disponibilidad de genomas completos ha sido la filogenia. Existente desde las teorías de

Darwin, la filogenia es el campo dentro de la biología dedicado al estudio de las relaciones evolutivas entre diferentes especies, siendo los árboles filogenéticos su instrumento más característico. Durante mucho tiempo, dichas relaciones se han venido estableciendo mediante la cuidadosa comparación de aquellos caracteres morfológicos que eran considerados homólogos entre organismos. Sin embargo, la mayoría de estos estudios han trasladado sus métodos hacia perspectivas más moleculares. La comparación de genes y proteínas permite un nivel de resolución mucho mayor que el aportado por los caracteres morfológicos, ya que, cada residuo (nucleótido o amino ácido que compone una secuencia) es considerado como un carácter independiente que puede ser comparado con sus homólogos en otras especies.

Este aumento de precisión conlleva, no obstante, un aumento de complejidad en los análisis. Las cadenas de nucleótidos y amino ácidos pueden llegar a ser extremadamente largas (hasta decenas de miles de posiciones), por lo que, hoy en día, su estudio no puede ser concebido sin la ayuda de ordenadores y algoritmos informáticos. La estrecha relación entre la informática y los estudios evolutivos basados en datos genómicos ha contribuido en gran medida a la formación y desarrollo de un área conocida como Bioinformática, dedicada a la aplicación de métodos computacionales a problemas biológicos. Así, la inferencia de filogenias moleculares se sustenta sobre una serie análisis computacionales encadenados que componen un protocolo de trabajo bastante estandarizado:

En primer lugar, es necesario contar con un grupo de secuencias relacionadas evolutivamente sobre las cuales deseamos conocer sus relaciones de parentesco. A continuación, estas secuencias han de ser alineadas para establecer las correspondencias de homología entre cada uno de sus residuos. El resultado de este paso es una matriz, llamada alineamiento múltiple de secuencias, en las que cada nucleótido o amino ácido es colocado justo encima de su correspondiente homólogo en el resto de secuencias. Cuando el residuo homólogo no existe en alguna de las secuencias, un hueco es insertado en su lugar. Finalmente, para deducir las relaciones entre las cadenas, es necesario aplicar un método de inferencia filogenética que, mediante la comparación de todos las posiciones del alineamiento, deduce la distancia evolutiva entre todas las secuencias. El resultado es típicamente representado en forma de un árbol dicotómico conocido como árbol filogenético. En la actualidad, existen bastantes algoritmos para alinear automáticamente grupos extensos de secuencias, y un amplio abanico de métodos para inferir árboles filogenéticos. En general, estos últimos se dividen en tres grandes grupos: los basados en parsimonia (poco usados), los basados en métodos de distancia (muy extendidos pero poco precisos) y los estadísticos (los más precisos pero muy costosos computacionalmente)

A.2 Homología, Paralogía y Ortología

Un aspecto importante a tener en cuenta en cualquier estudio filogenético basado en secuencias es que el análisis sólo tiene sentido en el contexto de secuencias relacionadas evolutivamente. Dicha condición es formalmente expresada mediante

el termino "homología" . La mayor parte de los métodos para detectar secuencias potencialmente homólogas se basan en test estadísticos que evalúan el grado de similitud entre sus residuos.

La relación de homología entre dos secuencias puede ser debida a dos tipos de procesos: (i) un proceso de especiación que provocó la divergencia independiente de dos genomas (y de todas sus secuencias) (ii) un proceso de duplicación génica que dio lugar a dos copias de la misma secuencia que divergieron como genes independientes. En el primer caso, la relación de homología es conocida como ortología; mientras que para el segundo se utiliza el término paralogía. La distinción entre ambos tipos cobra especial importancia para estudiar procesos concretos como la aparición de nuevos genes y funciones, o para establecer las correspondencias entre genes de diferentes organismos.

A.3 Filogenómica, filomas y variabilidad filogenética

La revolución ocurrida en el campo de la secuenciación genómica ha generado una explosión de datos. Sólo en los últimos 10 años, el número de especies totalmente secuencias ha pasado de una decena a casi el millar. De acuerdo a la base de datos GOLD (Genome OnLine Database) , cerca de otros 2000 proyectos están hoy en marcha. Tal cantidad de información ha generado un conjunto de áreas conocido como 'omicas'. En su mayoría, las ómicas son variantes de campos ya existentes que han extendido su visión hacia una perspectiva más amplia (genómica) por medio de la integración de toda la información concerniente a un organismo. Así, la filogenómica, área en dónde se encuadra esta tesis, es entendida como la variante de la filogenia que basa sus estudios en el uso de grandes cantidades de secuencias.

Los datos obtenidos por medio de las ómicas son a menudo nombrados mediante el sufijo 'oma'. Así, el conjunto de genes de un organismo es su genoma; el conjunto de proteínas es su proteoma; el conjunto de interacciones entre la proteínas es su interactoma; y el conjunto de filogenias de los genes de una especie es llamado **filoma**. El concepto de filoma constituye el tema central de esta tesis. Aunque su reconstrucción y análisis representan un valioso recurso para el estudio de muchos procesos evolutivos, su uso ha estado tradicionalmente impedido por un alto coste computacional.

Una de sus aplicaciones más directas puede encontrarse en el campo de la taxonomía. Dado que los filomas contienen la historia evolutiva de todos los genes de un organismo, éstos podrían determinar más precisamente las relaciones entre especies. Sin embargo, se ha comprobado en numerosas ocasiones que el análisis de dos secuencias de un mismo organismo puede llevar a resultados incompatibles entre sí. Es por ello que surja la siguiente pregunta: ¿podría el uso de filomas completos aumentar la señal filogenética? Esta posibilidad ha sido cuestionada en varias ocasiones, y si bien un aumento en la cantidad de información puede contribuir a aumentar las señal del estudio, éste también puede aumentar el ruido,

y por tanto, la incongruencia entre resultados. A lo largo del Capítulo 5 de esta tesis, se explora la posibilidad de usar los filomas como instrumento de medición de variabilidad filogenética.

A.4 Duplicación génica y evolución de genomas

La duplicación génica constituye el sustrato principal por el que la evolución crea y modela nuevas funciones en los genomas . Esto se explica debido a que la eventual duplicación de un gen da lugar a un estado de redundancia genética que puede ser "aprovechado" para explorar las ventajas de cambios genéticos sin necesidad de alterar la funcionalidad original. Cuando estos cambios aportan alguna ventaja al individuo, el gen duplicado se retiene en el genoma y es heredado por las siguientes generaciones. Por esta razón, el estudio del proceso de duplicación génica representa un excelente contexto para entender la divergencia entre genomas completos, y como éstos adquieren nuevas características y funciones.

Existen dos hipótesis principales para explicar la retención de genes duplicados en los genomas: Los llamados modelos de sub- y neo-funcionalización. El primero de ellos postula que cuando un gen se duplica, sendas copias pueden evolucionar hasta adquirir patrones funcionales complementarios, situación en la que la función original sólo puede ser llevada a cabo mediante la combinación de ambas copias. Por el contrario, la neo-funcionalización predice que mientras una de las copias retiene la función original, su duplicado es libre de evolucionar hasta adquirir un nuevo rol en el genoma. En cualquiera de los dos casos, la retención de ambas copias se vería favorecida a lo largo de la evolución.

A.5 Expresión divergente entre genes duplicados

La expresión génica puede ser perfectamente entendida como un aspecto funcional, ya que ésta define, por ejemplo, el patrón temporal y espacial en el que un gen desarrolla sus acciones. Es por ello que los modelos de sub- y neo-funcionalización puedan ser aplicados para entender la dinámica existente en los patrones de expresión.

En el capítulo 7 de esta tesis se aborda, concretamente, la divergencia en el patrón de expresión por tejidos que tiene lugar entre genes duplicados.

A.6 Reconstrucción del filoma humano

Al igual que otros resultados de carácter genómico (interactomas, metabolomas, proteomas, etc.), el filoma humano constituye una importante fuente de información para la caracterización de nuestro genoma. En el Capítulo 5 hemos abordado su reconstrucción a gran escala mediante el uso de métodos filogenéticos de gran precisión. El análisis ha sido llevado a cabo en el contexto de 38

especies eucariotas que cubren un amplio rango de organismos modelo en investigación. El procedimiento usado para construir las filogenias incluye pasos como el preprocesado de alineamientos, la evaluación de distintos modelos evolutivos, la estimación de sitios invariantes y la inferencia mediante métodos estadísticos (máxima verosimilitud e inferencia bayesiana). En suma, el procedimiento global ha sido diseñado para reproducir de forma automática la secuencia de acciones que habría seguido un filogenetista para realizar los análisis de la forma más precisa posible. Estos métodos fueron aplicados a los más de 20.000 genes que componen el genoma humano y se obtuvieron un total de 157,233 filogenias (evaluando diferentes modelos) y 21,278 alineamientos múltiples. La ejecución de todos los cálculos llevó cerca de 2 meses de computación continuada en varios súper-ordenadores (200 CPU paralelas).

A.7 Estudio de variabilidad filogenética

Se realizaron dos análisis topológicos sobre la totalidad de árboles obtenidos en el filoma humano:

En primer lugar, el conjunto de filogenias fueron evaluadas para descubrir posibles irrupciones en las relaciones taxonómicas más claramente establecidas y que pudieran, por tanto, indicar casos de transferencia horizontal de genes. Aunque un pequeño numero de estos casos pudo ser detectado, ninguno de ellos presentaba un soporte estadístico suficiente para ser consistente.

Por otro lado, la aplicación de la filogenómica al estudio taxonómico permite obtener una visión general del tipo y grado de señal evolutiva presente en todos los genes de un organismo. Ya que las relaciones entre algunas de las especies incluidas en el filoma humano no están aún totalmente resueltas, su estudio filogenómico podría proporcionar más información al respecto. Con esta finalidad, se llevó a cabo un segundo estudio topológico centrado en tres aspectos específicos y ampliamente discutidos en la literatura : la posición relativa de nemátodos, cordados y artrópodos; las relaciones entre roedores, primates y laurasiaterios; y la agrupación de opistocontos y amebozoos. El resultado, en resumen, mostró una elevada variabilidad, más incluso de la esperada, el los tres escenarios. Aunque en todos los casos una de las hipótesis siempre se mostró sobrerepresentada, las diferencias frente al resto no fueron suficientemente significativas. Así, lejos de clarificar las relaciones entre los escenarios evaluados, este resultado pone en evidencia la falta de potencia (incluso a escala genómica) de los métodos actuales para dilucidar algunas de la problemáticas actuales en taxonomía.

A.8 Filomas aplicados a la predicción de ortología

Dada la gran cantidad de tiempo y recursos computacionales asociados a la reconstrucción filogenética, las predicciones de ortología (un concepto puramente evolutivo) entre genes de diferentes especies se realiza, normalmente, mediante

enfoques menos costosos como los basados en similitud de secuencia. Aunque estos métodos son más rápidos y bastante fiables, han mostrado también ciertos inconvenientes a la hora de descubrir relaciones muy intrincadas entre genes.

Con el fin de aprovechar la información evolutiva presente en el filoma humano, en esta tesis se ha presentado un método alternativo de predicción de ortología basado en la interpretación automática de árboles filogenéticos (ver método en 11.1). Brevemente, nuestro algoritmo analiza la topología de las filogenias e infiere los eventos evolutivos basándose en el nivel de solapamiento entre las especies agrupadas por ramas hermanas. Nuestros resultados han arrojado un significativo incremento de precisión y poder predictivo en comparación con el resto de procedimientos (ver 5.3).

A.9 Identificación y datado de eventos de duplicación génica

La posibilidad de conocer la edad del evento de duplicación que originó a dos genes actuales supone una gran ventaja en el estudio de la evolución de los genomas. No obstante, el momento en el que una duplicación génica tuvo lugar sólo puede ser inferido mediante estudios indirectos. Por ejemplo, el método más extendido para realizar este tipo de análisis se basa en el grado de divergencia sinónima entre dos secuencias. La divergencia sinónima es entendida como la tasa de mutaciones sinónimas (dS) (cambios en codones sin efecto sobre el amino ácido que codifican) acumuladas entre dos secuencias. Ya que los cambios sinónimos no afectan al nivel de adaptación del individuo, generalmente se asume que éstos se acumulan de forma linear en el tiempo. Así, conociendo la tasa de cambios no sinónimos entre dos genes, podemos obtener una medida relativa de su tiempo de divergencia. Sin embargo, el uso de dS para establecer el tiempo evolutivo entre secuencias muy lejanas presenta ciertos inconvenientes, ya que esta tasa se satura muy fácilmente y puede llegar a ser muy variable.

En el Capítulo 6 se ha evaluado la posibilidad de usar el análisis de filogenias para obtener una estima más precisa de la edad de los eventos de duplicación génica. Para realizar la comparación, se calculó el dS de 250 genes parálogos de levadura cuyo origen está establecido en un mismo punto evolutivo: la duplicación del genoma completo ocurrida en el ancestro de esta especie hace unos 100Ma y avalada por estudios de sintenia. Los resultados mostraron que, lejos de obtener una distribución homogénea de tasas, los 250 genes parálogos arrojaron una gran variabilidad de valores dS. Esto puede dar idea de los inconvenientes derivados de la presunción de linealidad entre dS y el tiempo de divergencia. Por el contrario, nuestras predicciones de tiempo derivadas del análisis topológico de filogenias mostraron una distribución menos dispersa y correctamente centrada en el momento estimado para la duplicación masiva que dio lugar a los 250 parálogos (Figura 6.2).

La misma variabilidad de dS se observó en un análisis similar realizado sobre los genes duplicados humanos que previamente se habían asignado, mediante

métodos filogenéticos, a diferentes periodos evolutivos.

A.10 Evolución de los perfiles de expresión entre genes duplicados

Los cambios de expresión constituyen un tipo de variación funcional muy común entre genes. Estos cambios pueden ser debidos a modificaciones en la cantidad de expresión o en su patrón espacio-temporal. Dado que los genes duplicados son especialmente susceptibles a fijar cambios de función, su estudio en el contexto de las variaciones transcripcionales resulta muy útil para entender tanto la dinámica general de retención de parálogos como la evolución de los patrones de expresión.

En el Capítulo 6 se han abordado varios análisis sobre el caso específico de divergencia en expresión que conlleva cambios en el patrón espacial de tejidos. En primer lugar, se comparó el nivel de expresión complementaria (número de tejidos en los que se expresa un gen pero no otro) entre pares de genes parálogos y ortólogos. Aunque el nivel observado fue bastante elevado en ambos casos, los genes parálogos mostraron una mayor y significativa predisposición a adquirir (o retener) patrones de expresión complementarios. A diferencia de otros estudios previos, no se observó correlación linear entre la complementariedad de expresión y la edad de los duplicados.

Finalmente, y con el fin de estimar el momento relativo en el dos genes duplicados podrían fijar patrones complementarios de expresión, se realizó un análisis en el que se calculó la cantidad y tipo de complementariedad entre los parálogos humanos y sus homólogos en ratón. El paralelismo mostrado entre ambas observaciones, cuando existía complementariedad entre genes humanos también existía en ratón y la complementariedad implicaba a los mismos tejidos, sugiere que la complementariedad tuvo que ser adquirida a priori del evento de especiación entre la especie humana y ratón. Sin embargo, y de forma interesante, fue posible encontrar algunos ejemplos en los que la complementariedad entre parálogos era debida a tejidos mucho más modernos que la propia duplicación.

A.11 PhylomeDB y ETE, dos recursos públicos para el análisis filogenómico

La utilidad de los datos genómicos y la posibilidad de que éstos sean extendidos o aplicados a diferentes problemáticas depende, en gran medida, de su disponibilidad. Por esa razón, uno de los objetivos de esta tesis ha sido promover el uso de los datos generados y facilitar el acceso a los métodos desarrollados. En el capítulo 8 se describen dos recursos filogenómicos que han sido puestos a disposición de la comunidad científica: PhylomeDB y ETE (Environment for Tree Exploration).

PhylomeDB es una base de datos relacional que aloja las filogenias y alineamientos de secuencias derivados de la reconstrucción de filomas completos. La

base de datos puede ser consultada vía web mediante un interfaz muy intuitivo que permite tanto la descarga directa de información como su visualización interactiva. Los árboles filogenéticos resultantes de todos los análisis (examinando, por ejemplo, el ajuste de diferentes modelos evolutivos) y dos versiones (cruda y procesada) de los alineamientos múltiples de secuencias están disponibles para todos los filomas registrados. Además, las predicciones de ortología y paralogía extraídas del análisis de cada filoma pueden ser visualizadas sobre los mismos árboles filogenéticos que se usaron para su identificación.

ETE es un conjunto de librerías de programación que pueden ser usadas para explotar los métodos y análisis desarrollados en esta tesis. La funcionalidad básica de ETE constituye un entorno de trabajo que facilita el manejo, la manipulación y la visualización de árboles filogenéticos. Mediante una serie de paquetes extra, ETE implementa además un buen número de algoritmos que cubren aspectos como la detección de eventos evolutivos, el datado topológico de los mismos, el enraizado de las filogenias o su análisis visual. Además, ETE proporciona acceso remoto a la base de datos phylomeDB, por lo que puede ser usado para explotar de forma automática los datos precalculados existentes en dicha base de datos. En la actualidad el paquete ETE está siendo utilizado por proyectos como GEPAS o Phylemon y puede ser descargado de `http://bioinfo.cipf.es/downloads/ete`.

A.12 Discusión global

A lo largo de esta tesis se han tratado cuatro temas principales: (i) La reconstrucción completa del Filoma Humano mediante el uso de métodos filogenéticos de alta precisión, (ii) el uso de filomas como herramienta para la esclarecer las relaciones evolutivas entre diferentes especies y las relaciones de ortología y paralogía entre sus genes, (iii) el estudio filogenético del proceso de duplicación génica y sus implicaciones en la evolución de los perfiles de expresión entre genes parálogos, (iv) y la construcción de una base de datos pública y varias herramientas bioinformáticas que recopilan los resultados y métodos obtenidos en esta tesis.

Reconstrucción de filomas

En cuanto al primero de los objetivos, hemos demostrado que el uso de métodos filogenéticos de alta precisión es hoy en día posible a nivel genómico, si bien a un elevado coste computacional. La reconstrucción del filoma humano llevó más de dos meses de computación continuada en varios súper ordenadores, lo que equivale aproximadamente a 2 años en una sola maquina. Esto supone el mayor esfuerzo realizado hasta la fecha para la reconstrucción de la historia evolutiva de todos los genes humanos. Además, la metodología desarrollada se está mostrando muy útil en su aplicación a otros contextos, especies y escenarios evolutivos.

Por otro lado, la visión filogenética global aportada por el filoma humano nos ha permitido comprobar la existencia de una elevada variabilidad evolutiva entre diferentes genes. Aunque esta variabilidad parece afectar principalmente a

relaciones entre grupos taxonómicos específicos (coincidiendo con los casos más debatidos), nuestros resultados sugieren que ha de prestarse especial atención al grupo de genes que es usado para extraer conclusiones genómicas, ya que diferentes genes podrían estar sesgados hacia hipótesis distintas.

Predicción de eventos de duplicación y especiación.

La posibilidad de automatizar todos los pasos necesarios para reconstruir filomas completos, así como para derivar automáticamente algunos de los análisis, dota al conjunto de métodos presentado en esta tesis de una excelente flexibilidad. Por ejemplo, su uso nos ha permitido obtener, de forma no supervisada, el catálogo completo de ortólogos y parálogos de todo el genoma humano en un contexto de otras 38 especies. Estas predicciones han mostrado, además, una mejora sustancial en términos de sensitividad y poder predictivo con respecto a métodos previos.

Nuestro punto de vista es que este tipo de metodología resulta muy útil, no solo para mejorar las predicciones de ortología ya existentes, sino para ser aplicado de forma automática a genomas de organismos no modelo o recién secuenciados.

Un modelo mixto de evolución de los perfiles de expresión entre genes duplicados

Otro de los aspectos más relevantes tratados en esta tesis es la dinámica observada en los patrones de expresión de genes duplicados. Para medir los cambios de expresión en entre los genes, la mayoría de estudios realizados hasta la fecha utilizan el nivel de correlación entre los perfiles de expresión de dos genes duplicados como medida de divergencia. Muchos de estos trabajos han observado, además, que existe una relación linear inversa entre el nivel de divergencia de expresión y tiempo, es decir, que los genes que han divergido durante más tiempo acumulan una mayor diferencia en sus patrones de expresión. Esta observación sugiere que las variaciones de expresión podrían estar sujetas a un modelo de evolución neutra , por el que la mayor parte de los cambios de expresión serían evolutivamente inocuos y tenderían a acumularse con el tiempo. Ya que esto no es lo que inicialmente esperaríamos bajo los modelos de sub- y neo-funcionalización, nosotros hemos sugerido (Capítulo 7) que diferentes tipos de expresión divergente podrían estar actuando al mismo tiempo.

Para probar esta idea, los análisis presentados en el Capítulo 6 han tratado de aislar un tipo concreto de expresión divergente que estaría, en principio, más claramente asociado a la retención de duplicados: los cambios en el registro de tejidos en el que se expresan un grupo de genes parálogos. Al contrario de lo observado en trabajos previos, nuestros resultados no muestran una correlación significativa entre el índice de complementariedad de expresión por tejidos y el tiempo de divergencia, apoyando así las predicciones de los modelos de sub- y neo-funcionalización. Además, las diferencias significativas observadas entre la

expresión complementaria de parálogos y ortólogos sugieren que una fuerza selectiva está actuando en favor de los cambios de expresión en genes parálogos. Por último, el relativamente corto espacio de tiempo en el que los parálogos tienden a adquirir patrones de expresión complementarios refuerza la idea de un modelo de retención de genes duplicados.

Por otro lado, el alto nivel de expresión complementaria observado entre genes ortólogos lleva a pensar que otro tipo de variaciones, esta vez neutrales, podría estar actuando al mismo tiempo. En este sentido, el hecho de haber encontrado genes parálogos especializados en tejidos mucho más modernos que el propio origen de su evento de duplicación, sugiere un cierto grado de "reciclaje" de los patrones de expresión génica.

Como conclusión, nosotros proponemos un modelo mixto para la evolución de los perfiles de expresión, en el que cierto grado de variación sería mayoritariamente neutro, y otra parte estaría bajo presión selectiva.

A.13 Conclusiones

1. Mediante la reconstrucción del Filoma Humano, hemos demostrado que es actualmente posible reconstruir filomas completos aplicando los métodos filogenéticos más potentes hasta la fecha. La metodología presentada se sustenta sobre una línea de trabajo totalmente automatizada que abre camino hacia su aplicación a otras especies o escenarios.

2. El análisis topológico del filoma humano ha reforzado la vieja idea de que diferentes genes de un mismo organismo pueden apuntar hacia historias evolutivas contradictorias, incluso en ausencia de transferencia horizontal de genes. Nosotros compartimos la visión de otros trabajos previos de que estos problemas sólo pueden ser resueltos mediante mejoras en los métodos y modelos de análisis.

3. Hemos implementado un algoritmo capaz de extraer automáticamente la información más relevante de grandes cantidades de árboles filogenéticos. Este método ha mostrado mejoras en las predicciones de ortología y paralogía cuando es aplicado a casos en los que el resto de métodos se comportan irregularmente. Proponemos, por ello, el uso de filogenias de alta calidad para establecer dichas relaciones entre nuevas especies, o para mejorar las predicciones ya existentes.

4. Hemos discutido los inconvenientes de usar las tasas de cambios sinónimos (dS) como predictores del tiempo de divergencia entre genes duplicados. Como alternativa, proponemos el uso de filogenias para detectar y datar los eventos de duplicación.

5. Las observaciones obtenidas del análisis evolutivo de los patrones de expresión entre genes duplicados nos ha llevado a proponer un modelo mixto de evolución, en el que los variaciones de expresión pueden ser tanto neutras como favorables para la retención de parálogos.

6. Hemos desarrollado una base de datos relacional que aloja toda la información obtenida del filoma humano, así como de nuevos filomas. Todos los ár-

boles filogenéticos, alineamientos de secuencias y predicciones de ortología están disponibles públicamente en http://phylomedb.bioinfo.cipf.es

7. Hemos desarrollado ETE (Environment for Tree Exploration), un paquete de herramientas bioinformáticas que, además de permitir el manejo y visualización de árboles filogenéticos, implementa los algoritmos de análisis empleados en esta tesis y un módulo de acceso a la base de datos phylomeDB.

Apéndice B

List of publications

1. **Huerta-Cepas J**, Dopazo H, Dopazo J, Gabaldón T. *The human phylome.* Genome Biol. 2007;8(6):R109.
2. **Huerta-Cepas J**, Bueno A, Dopazo J, Gabaldón T. *PhylomeDB: a database for genome-wide collections of gene phylogenies.* Nucleic Acids Res. 2008 Jan;36(Database issue):D491-6.
3. **Huerta-Cepas J**, Gabaldón T, Dopazo J. *ETE: A python programming Environment for Tree Exploration.* (submitted)
4. **Huerta-Cepas J**, Dopazo J, Huynen MA, Gabaldón T. *Evolutionary dating of human duplicated genes and their tissue expression divergence.* (manuscript in preparation)
5. **Huerta-Cepas J**, Dopazo J, Gabaldón T. *Evolutionary dating of duplication events: topological dating as an alternative to the use of synonymous substitution rates.* (manuscript in preparation)
6. Tarraga J, Medina I, Arbiza L, **Huerta-Cepas J**, Gabaldón T, Dopazo J, Dopazo H. *Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics.* Nucleic Acids Res. 2007 Jul;35(Web Server issue):W38-42.
7. Schlüter A, Fourcade S, Domènech-Estévez E, Gabaldón T, **Huerta-Cepas J**, Berthommier G, Ripp R, Wanders RJ, Poch O, Pujol A. *PeroxisomeDB: a database for the peroxisomal proteome, functional genomics and disease.* Nucleic Acids Res. 2007 Jan;35(Database issue):D815-22.
8. Montaner D, Tárraga J, **Huerta-Cepas J**, Burguet J, Vaquerizas JM, Conde L, Minguez P, Vera J, Mukherjee S, Valls J, Pujana MA, Alloza E, Herrero J, Al-Shahrour F, Dopazo J. *Next station in microarray data analysis: GEPAS.* Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W486-91.
9. Tárraga J, Medina I, Carbonell J, **Huerta-Cepas J**, Minguez P, Alloza E, Al-Shahrour F, Vegas-Azcárate S, Goetz S, Escobar P, Garcia-Garcia F, Conesa A, Montaner D, Dopazo J. *GEPAS, a web-based tool for microarray data analysis and interpretation.* Nucleic Acids Res. 2008 Jul 1;36(Web Server issue):W308-14.
10. Al-Shahrour F, Arbiza L, Dopazo H, **Huerta-Cepas J**, Mínguez P, Montaner D, Dopazo J. *From genes to functional classes in the study of biological systems.* BMC Bioinformatics. 2007 Apr 3;8:114.

11. Hernández P, **Huerta-Cepas J**, Montaner D, Al-Shahrour F, Valls J, Gómez L, Capellá G, Dopazo J, Pujana MA. *Evidence for systems-level molecular mechanisms of tumorigenesis.* BMC Genomics. 2007 Jun 20;8:185.

12. BioMoby Consortium, Wilkinson MD, Senger M, Kawas E, Bruskiewich R, Gouzy J, Noirot C, Bardou P, Ng A, Haase D, Saiz Ede A, Wang D, Gibbons F, Gordon PM, Sensen CW, Carrasco JM, Fernández JM, Shen L, Links M, Ng M, Opushneva N, Neerincx PB, Leunissen JA, Ernst R, Twigger S, Usadel B, Good B, Wong Y, Stein L, Crosby W, Karlsson J, Royo R, Párraga I, Ramírez S, Gelpi JL, Trelles O, Pisano DG, Jimenez N, Kerhornou A, Rosset R, Zamacola L, Tarraga J, **Huerta-Cepas J**, Carazo JM, Dopazo J, Guigo R, Navarro A, Orozco M, Valencia A, Claros MG, Pérez AJ, Aldana J, Rojano MM, Fernandez-Santa Cruz R, Navas I, Schiltz G, Farmer A, Gessler D, Schoof H, Groscurth A. *Abstract Interoperability with Moby 1.0–it's better than sharing your toothbrush!* Brief Bioinform. 2008 May;9(3):220-31.

Bibliografía

Abhiman, S. and Sonnhammer, E. L. L. (2005). Funshift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Research*, **33**(Database issue), D197–200. PMID: 15608176.

Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial dna. *Journal of Molecular Evolution*, **42**(4), 459–68. PMID: 8642615.

Adams, M. and Fleischmann, R. (1995). Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, **269**(5223), 496–512.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., and Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, **387**(6632), 489–93. PMID: 9168109.

Akaike, H. (1973). Information theory and extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory: 1973; Budapest, Hungary. Edited by: Institute of Electrical and Electronics Engineers. Piscataway*, pages 267–281.

Al-Shahrour, F., Minguez, P., Tárraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J., and Dopazo, J. (2006). Babelomics: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research*, **34**(Web Server issue), W472–6. PMID: 16845052.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–10. PMID: 2231712.

Alvarez, N., Benrey, B., Hossaert-McKey, M., Grill, A., McKey, D., and Galtier, N. (2006). Phylogeographic support for horizontal gene transfer involving sympatric bruchid species. *Biology Direct*, **1**, 21. PMID: 16872524.

Andersson, J. O., Sjögren, A. M., Davis, L. A. M., Embley, T. M., and Roger, A. J. (2003). Phylogenetic analyses of diplomonad genes reveal frequent lateral gene transfers affecting eukaryotes. *Current Biology: CB*, **13**(2), 94–104. PMID: 12546782.

Arvestad, L., Berglund, A.-C., Lagergren, J., and Sennblad, B. (2003). Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics (Oxford, England)*, **19 Suppl 1**, i7–15.

Bailey, J. A. and Eichler, E. E. (2006). Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews. Genetics*, **7**(7), 552–64. PMID: 16770338.

Benner, S. A. (2003). Interpretive proteomics–finding biological meaning in genome and proteome databases. *Advances in Enzyme Regulation*, **43**, 271–359. PMID: 12791396.

Berglund-Sonnhammer, A.-C., Steffansson, P., Betts, M. J., and Liberles, D. A. (2006). Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *Journal of Molecular Evolution*, **63**(2), 240–50. PMID: 16830091.

Bergthorsson, U., Adams, K. L., Thomason, B., and Palmer, J. D. (2003). Widespread horizontal transfer of mitochondrial genes in flowering plants. *Nature*, **424**(6945), 197–201. PMID: 12853958.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Gräf, S., Hammond, M., Herrero, J., Howe, K., Iyer, V., Jekosch, K., Kähäri, A., Kasprzyk, A., Keefe, D., Kokocinski, F., Kulesha, E., London, D., Longden, I., Melsopp, C., Meidl, P., Overduin, B., Parker, A., Proctor, G., Prlic, A., Rae, M., Rios,

D., Redmond, S., Schuster, M., Sealy, I., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Stabenau, A., Stalker, J., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., and Hubbard, T. J. P. (2006). Ensembl 2006. *Nucleic Acids Research*, **34**(Database issue), D556–61. PMID: 16381931.

Blackstone, N. W. and Green, D. R. (1999). The evolution of a mechanism of cell suicide. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **21**(1), 84–8. PMID: 10070258.

Blair, J. E. and Hedges, S. B. (2005). Molecular phylogeny and divergence times of deuterostome animals. *Molecular Biology and Evolution*, **22**(11), 2275–84. PMID: 16049193.

Blomme, T., Vandepoele, K., Bodt, S. D., Simillion, C., Maere, S., and de Peer, Y. V. (2006). The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biology*, **7**(5), R43. PMID: 16723033.

Bromham, L. (2002). The human zoo: endogenous retroviruses in the human genome. *Trends Ecol Evol*, (17), 160.

Bruno, W. J. and Halpern, A. L. (1999). Topological bias and inconsistency of maximum likelihood using wrong models. *Molecular Biology and Evolution*, **16**(4), 564–6. PMID: 10331281.

Buckley, T. R. and Cunningham, C. W. (2002). The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Molecular Biology and Evolution*, **19**(4), 394–405. PMID: 11919280.

Byrne, K. P. and Wolfe, K. H. (2005). The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Research*, **15**(10), 1456–61. PMID: 16169922.

Castresana, J. (2007). Topological variation in single-gene phylogenetic trees. *Genome Biology*, **8**(6), 216. PMC2394742.

Cavalier-Smith, T. (2002). The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. *International Journal of Systematic and Evolutionary Microbiology*, **52**(Pt 2), 297–354. PMID: 11931142.

Clamp, M., Cuff, J., Searle, S. M., and Barton, G. J. (2004). The jalview java alignment editor. *Bioinformatics*, **20**(3), 426–427.

Consortium, I. H. G. S. (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Dagan, T. and Martin, W. (2006). The tree of one percent. *Genome Biology*, **7**(10), 118. PMID: 17081279.

Delsuc, F., Brinkmann, H., and Philippe, H. (2005). Phylogenomics and the reconstruction of the tree of life. *Nature Reviews. Genetics*, **6**(5), 361–75. PMID: 15861208.

Dopazo, H. and Dopazo, J. (2005). Genome-scale evidence of the nematode-arthropod clade. *Genome Biology*, **6**(5), R41. PMID: 15892869.

Dufayard, J.-F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière, G. (2005). Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics (Oxford, England)*, **21**(11), 2596–603. PMID: 15713731.

Duret, L., Mouchiroud, D., and Gouy, M. (1994). Hovergen: a database of homologous vertebrate genes. *Nucleic Acids Research*, **22**(12), 2360–5. PMID: 8036164.

Eddy, S. R. (1995). Multiple alignment using hidden markov models. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **3**, 114–20. PMID: 7584426.

Edgar, R. C. (2004). Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113. PMID: 15318951.

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, **8**(3), 163–7. PMID: 9521918.

Eisen, J. A., Kaiser, D., and Myers, R. M. (1997). Gastrogenomic delights: a movable feast. *Nature Medicine*, **3**(10), 1076–8. PMID: 9334711.

Fisher, S. E. and Marcus, G. F. (2006). The eloquent ape: genes, brains and the evolution of language. *Nature Reviews. Genetics*, **7**(1), 9–20. PMID: 16369568.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, **19**(2), 99–113. PMID: 5449325.

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics: TIG*, **16**(5), 227–31. PMID: 10782117.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., and Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, **151**(4), 1531–1545.

Freilich, S., Massingham, T., Blanc, E., Goldovsky, L., and Thornton, J. M. (2006). Relating tissue specialization to the differentiation of expression of singleton and duplicate mouse proteins. *Genome biology*, **7**, R89. PMID: 17029626.

Gabaldón, T. and Huynen, M. A. (2003). Reconstruction of the proto-mitochondrial metabolism. *Science (New York, N.Y.)*, **301**(5633), 609. PMID: 12893934.

Gabaldón, T. and Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. *Cellular and molecular life sciences : CMLS*, **61**, 930–44.

Gabaldón, T. and Huynen, M. A. (2005). Lineage-specific gene loss following mitochondrial endosymbiosis and its potential for function prediction in eukaryotes. *Bioinformatics (Oxford, England)*, **21 Suppl 2**, ii144–50. PMID: 16204094.

Gabaldón, T., Rainey, D., and Huynen, M. A. (2005). Tracing the evolution of a large protein complex in the eukaryotes, nadh:ubiquinone oxidoreductase (complex i). *Journal of Molecular Biology*, **348**(4), 857–70. PMID: 15843018.

Gandhi, T. K. B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., Mishra, G., Nandakumar, K., Shen, B., Deshpande, N., Nayak, R., Sarker, M., Boeke, J. D., Parmigiani, G., Schultz, J., Bader, J. S., and Pandey, A. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature Genetics*, **38**(3), 285–93. PMID: 16501559.

Gascuel, O. (1997). Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**(7), 685–95. PMID: 9254330.

Goldsmith, M. R., Shimada, T., and Abe, H. (2005). The genetics and genomics of the silkworm, bombyx mori. *Annual Review of Entomology*, **50**, 71–100. PMID: 15355234.

Gu, Z., Nicolae, D., Lu, H. H.-S., and Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in genetics : TIG*, **18**, 609–13. PMID: 12446139.

Gu, Z., Rifkin, S. A., White, K. P., and Li, W.-H. (2004). Duplicate genes increase gene expression diversity within and between species. *Nature genetics*, **36**, 577–9. PMID: 15122255.

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704. PMID: 14530136.

Hallet, M., Lagergren, J., and Tofigh, A. (2004). Simultaneous identification of duplications and lateral transfers. *In Proceedings of the Eighth Annual International Conference on Research In Computational Molecular Biology*, pages 347–356.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**(22), 10915–9. PMID: 1438297.

Huerta-Cepas, J., Dopazo, H., Dopazo, J., and Gabaldon, T. (2007). The human phylome. *Genome Biology*, **8**(6), R109.

Hulsen, T., Huynen, M. A., de Vlieg, J., and Groenen, P. M. A. (2006). Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, **7**(4), R31. PMID: 16613613.

Huminiecki, L. and Wolfe, K. H. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research*, **14**, 1870–9. PMID: 15466287.

Humphery-Smith, I. (2004). A human proteome project with a beginning and an end. *Proteomics*, **4**(9), 2519–21. PMID: 15352225.

Huynen, M. A. and Bork, P. (1998). Measuring genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(11), 5849–56. PMID: 9600883.

Jeffroy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics: TIG*, **22**(4), 225–31. PMID: 16490279.

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences: CABIOS*, **8**(3), 275–82. PMID: 1633570.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., and Mclnerney, J. O. (2006). Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evolutionary Biology*, **6**, 29. PMID: 16563161.

Keeling, P. J., Burger, G., Durnford, D. G., Lang, B. F., Lee, R. W., Pearlman, R. E., Roger, A. J., and Gray, M. W. (2005). The tree of eukaryotes. *Trends in Ecology & Evolution (Personal Edition)*, **20**(12), 670–6. PMID: 16701456.

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, **188**(4184), 107–16. PMID: 1090005.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics*, **39**, 309–38. PMID: 16285863.

Kullberg, M., Nilsson, M. A., Arnason, U., Harley, E. H., and Janke, A. (2006). Housekeeping genes for phylogenetic analysis of eutherian relationships. *Molecular Biology and Evolution*, **23**(8), 1493–503. PMID: 16751257.

Kurland, C. G. (2005). What tangled web: barriers to rampant horizontal gene transfer. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, **27**(7), 741–7. PMID: 15954096.

Kurland, C. G., Canback, B., and Berg, O. G. (2003). Horizontal gene transfer: a critical view. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(17), 9658–62. PMID: 12902542.

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Hériché, J.-K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L., Wong, G. K.-S., Zheng, W., Dehal, P., Wang, J., and Durbin, R. (2006). Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, **34**(Database issue), D572–80. PMID: 16381935.

Li, L., Stoeckert, C. J., and Roos, D. S. (2003). Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, **13**(9).

Li, W.-H., Yang, J., and Gu, X. (2005). Expression divergence between duplicate genes. *Trends in genetics : TIG*, **21**, 602–7. PMID: 16140417.

Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. C. (2006). The genomes on line database (gold) v. 2: a monitor of genome projects worldwide. *Nucleic Acids Res*, **34**, D332–D334.

Lynch, M. and Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**, 459–73. PMID: 10629003.

Makova, K. D. and Li, W.-H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome research*, **13**, 1638–45. PMID: 12840042.

Marcet-Houben, M. and Gabaldón, T. (2007). The tree versus the forest: the fungal tree of life and the topological diversity within the yeast phylome. *(submitted)*.

Meselson, M. and Stahl, F. W. (1958). The replication of dna in escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, **44**(7), 671–82. PMID: 16590258.

Messing, J., Crea, R., and Seeburg, P. H. (1981). A system for shotgun dna sequencing. *Nucleic Acids Research*, **9**(2), 309–21. PMID: 6259625.

Meyer, A. (2003). Molecular evolution: Duplication, duplication. *Nature*, **421**(6918), 31–2. PMID: 12511940.

Misawa, K. and Janke, A. (2003). Revisiting the glires concept–phylogenetic analysis of nuclear sequences. *Molecular Phylogenetics and Evolution*, **28**(2), 320–7. PMID: 12878468.

Murphy, W. J., Pevzner, P. A., and O'Brien, S. J. (2004). Mammalian phylogenomics comes of age. *Trends in Genetics: TIG*, **20**(12), 631–9. PMID: 15522459.

Müller, T. and Vingron, M. (2000). Modeling amino acid replacement. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **7**(6), 761–76. PMID: 11382360.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48**(3), 443–53.

Nielsen, J. and Oliver, S. (2005). The next wave in metabolome analysis. *Trends in Biotechnology*, **23**(11), 544–6. PMID: 16154652.

Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, **302**(1), 205–17. PMID: 10964570.

O'Brien, K. P., Remm, M., and Sonnhammer, E. L. L. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*, **33**(Database issue), D476–80. PMID: 15608241.

Ohno, S. S. (1970). *Evolution by gene duplication*. [[Springer Science Business Media|Springer-Verlag]].

Ohta, T. (1995). Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution*, **40**(1), 56–63. PMID: 7714912.

Owen, R. (1848). *On the Archetype and Homologies of the Vertebrate Skeleton*.

Page, R. D. and Charleston, M. A. (1997). From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, **7**(2), 231–40. PMID: 9126565.

Panopoulou, G., Hennig, S., Groth, D., Krause, A., Poustka, A. J., Herwig, R., Vingron, M., and Lehrach, H. (2003). New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Research*, **13**(6A), 1056–66. PMID: 12799346.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering*, **14**(9), 609–14. PMID: 11707606.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, **85**(8), 2444–8. PMID: 3162770.

Philippe, H., Snell, E. A., Bapteste, E., Lopez, P., Holland, P. W. H., and Casane, D. (2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. *Molecular Biology and Evolution*, **21**(9), 1740–52. PMID: 15175415.

Pruess, M., Kersey, P., and Apweiler, R. (2005). The integr8 project–a resource for genomic and proteomic data. *In Silico Biology*, **5**(2), 179–85. PMID: 15972013.

Ramani, A. K. and Marcotte, E. M. (2003). Exploiting the co-evolution of interacting proteins to discover interaction specificity. *Journal of Molecular Biology*, **327**(1), 273–84. PMID: 12614624.

Ricard, G., McEwan, N. R., Dutilh, B. E., Jouany, J.-P., Macheboeuf, D., Mitsumori, M., McIntosh, F. M., Michalowski, T., Nagamine, T., Nelson, N., Newbold, C. J., Nsabimana, E., Takenaka, A., Thomas, N. A., Ushida, K., Hackstein, J. H. P., and Huynen, M. A. (2006). Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment. *BMC Genomics*, **7**, 22. PMID: 16472398.

Rokas, A. (2008). Genomics. lining up to avoid bias. *Science*, **319**(5862), 416–417.

Rokas, A. and Carroll, S. B. (2006). Bushes in the tree of life. *PLoS Biology*, **4**(11), e352. PMID: 17105342.

Ronquist, F. and Huelsenbeck, J. P. (2003). Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*, **19**(12), 1572–4. PMID: 12912839.

Roth, C., Betts, M. J., Steffansson, P., Saelensminde, G., and Liberles, D. A. (2005). The adaptive evolution database (taed): a phylogeny based tool for comparative genomics. *Nucleic Acids Research*, **33**(Database issue), D495–7. PMID: 15608245.

Roth, C., Rastogi, S., Arvestad, L., Dittmar, K., Light, S., Ekman, D., and Liberles, D. A. (2007). Evolution after gene duplication: models, mechanisms, sequences,

systems, and organisms. *Journal of Experimental Zoology. Part B. Molecular and Developmental Evolution*, **308**(1), 58–73. PMID: 16838295.

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. (2001). Microbial genes in the human genome: lateral transfer or gene loss? *Science (New York, N.Y.)*, **292**(5523), 1903–6. PMID: 11358996.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of Molecular Biology*, **94**(3), 441–8. PMID: 1100841.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12), 5463–7. PMID: 271968.

Schlüter, A., Fourcade, S., Domènech-Estévez, E., Gabaldón, T., Huerta-Cepas, J., Berthommier, G., Ripp, R., Wanders, R. J. A., Poch, O., and Pujol, A. (2006). Peroxisomedb: a database for the peroxisomal proteome, functional genomics and disease. *Nucleic Acids Res.*

Seoighe, C., Johnston, C. R., and Shields, D. C. (2003). Significantly different patterns of amino acid replacement after gene duplication as compared to after speciation. *Molecular Biology and Evolution*, **20**(4), 484–90. PMID: 12654935.

Sicheritz-Pontén, T. and Andersson, S. G. (2001). A phylogenomic approach to microbial evolution. *Nucleic Acids Research*, **29**(2), 545–52. PMID: 11139625.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**(1), 195–7. PMID: 7265238.

Stamatakis, A., Ludwig, T., and Meier, H. (2005). Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**(4), 456–463.

Suzuki, Y. and Sugano, S. (2006). Transcriptome analyses of human genes and applications for proteome analyses. *Current Protein & Peptide Science*, **7**(2), 147–63. PMID: 16611140.

Talavera, G. and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**(4), 564–77. PMID: 17654362.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science (New York, N.Y.)*, **278**(5338), 631–7. PMID: 9381173.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., Schwartz, M. S., Weber, R. J., Kent, W. J., Karolchik, D., Bruen, T. C., Bevan, R., Cutler, D. J., Schwartz, S., Elnitski, L., Idol, J. R., Prasad, A. B., Lee-Lin, S.-Q., Maduro, V. V. B., Summers, T. J., Portnoy, M. E., Dietrich, N. L., Akhter, N., Ayele, K., Benjamin, B., Cariaga, K., Brinkley, C. P., Brooks, S. Y., Granite, S., Guan, X., Gupta, J., Haghighi, P., Ho, S.-L., Huang, M. C., Karlins, E., Laric, P. L., Legaspi, R., Lim, M. J., Maduro, Q. L., Masiello, C. A., Mastrian, S. D., McCloskey, J. C., Pearson, R., Stantripop, S., Tiongson, E. E., Tran, J. T., Tsurgeon, C., Vogt, J. L., Walker, M. A., Wetherby, K. D., Wiggins, L. S., Young, A. C., Zhang, L.-H., Osoegawa, K., Zhu, B., Zhao, B., Shu, C. L., Jong, P. J. D., Lawrence, C. E., Smit, A. F., Chakravarti, A., Haussler, D., Green, P., Miller, W., and Green, E. D. (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, **424**(6950), 788–93. PMID: 12917688.

Tzika, A. C., Helaers, R., Van de Peer, Y., and Milinkovitch, M. C. (2008). Mantis: a phylogenetic framework for multi-species genome comparisons. *Bioinformatics*, **24**(2), 151–157.

Tárraga, J., Medina, I., Arbiza, L., Huerta-Cepas, J., Gabaldón, T., Dopazo, J., and Dopazo, H. (2007). Phylemon: a suite of web tools for molecular evolution, phylogenetics and phylogenomics. *Nucleic Acids Research*, **35**, W38–42. PMID: 17452346.

Tárraga, J., Medina, I., Carbonell, J., Huerta-Cepas, J., Minguez, P., Alloza, E., Al-Shahrour, F., Vegas-Azcárate, S., Goetz, S., Escobar, P., Garcia-Garcia, F., Conesa, A., Montaner, D., and Dopazo, J. (2008). Gepas, a web-based tool for microarray data

analysis and interpretation. *Nucleic Acids Research*, **36**, W308–14. PMID: 18508806.

van Noort, V., Snel, B., and Huynen, M. A. (2003). Predicting gene function by conserved co-expression. *Trends in Genetics: TIG*, **19**(5), 238–42. PMID: 12711213.

Vogel, C. and Chothia, C. (2006). Protein family expansions and biological complexity. *PLoS Computational Biology*, **2**(5), e48. PMID: 16733546.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**(4356), 737–8. PMID: 13054692.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular Biology and Evolution*, **18**(5), 691–9. PMID: 11319253.

Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2004). Coelomata and not ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Research*, **14**(1), 29–36. PMID: 14707168.

Wong, K. M., Suchard, M. A., and Huelsenbeck, J. P. (2008). Alignment uncertainty and genomic analysis. *Science*, **319**(5862), 473–476.

Yang, Z. (1997). Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, **13**(5), 555–556.

Zhang, J. (2000). Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution*, **50**(1), 56–68. PMID: 10654260.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins. *Evolving genes and proteins. Academic Press, New York*, **97**.